Differential geometric least angle regression

Augugliaro, Luigi; Mineo, Angelo M.; Wit, Ernst C.

Published in:
Journal of the Royal Statistical Society. Series B: Statistical Methodology

DOI:
10.1111/rssb.12000

Publication date:
2013

Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2013)
75, *Part* 3, *pp.* 471–498

Differential geometric least angle regression: a differential geometric approach to sparse generalized linear models

Luigi Augugliaro and Angelo M. Mineo

University of Palermo, Italy

and Ernst C. Wit

University of Groningen, The Netherlands

[Received February 2010. Final revision August 2012]

Summary. Sparsity is an essential feature of many contemporary data problems. Remote sensing, various forms of automated screening and other high throughput measurement devices collect a large amount of information, typically about few independent statistical subjects or units. In certain cases it is reasonable to assume that the underlying process generating the data is itself sparse, in the sense that only a few of the measured variables are involved in the process. We propose an explicit method of monotonically decreasing sparsity for outcomes that can be modelled by an exponential family. In our approach we generalize the equiangular condition in a generalized linear model. Although the geometry involves the Fisher information in a way that is not obvious in the simple regression setting, the equiangular condition turns out to be equivalent with an intuitive condition imposed on the Rao score test statistics. In certain special cases the method can be tweaked to obtain L_1-penalized generalized linear model solution paths, but the method itself defines sparsity more directly. Although the computation of the solution paths is not trivial, the method compares favourably with other path following algorithms.

Keywords: Covariance penalty theory; Differential geometry; Generalized degrees of freedom; Generalized linear models; Information geometry; Least angle regression; Path following algorithm; Sparse models; Variable selection

1. Introduction

This paper deals with the problem of how to study the sparse structure of a generalized linear model (GLM) (McCullagh and Nelder, 1989). Modern statistical methods developed to cope with this problem are typically based on using a penalized objective function for estimating a solution curve embedded in the parameter space and then finding the point on that curve that represents the best compromise between sparsity and predictive behaviour of the model. Some of the most important examples are the L_1-penalty function that was originally proposed by Tibshirani (1996) for linear regression models, the smoothly clipped absolute deviation method that was proposed by Fan and Li (2001) and the Dantzig selector that was proposed by Candes and Tao (2007) and extended to GLMs in James and Radchenko (2009), among others.

In general the structure of the exponential family combined with the penalty function imposed to obtain sparsity of the GLM results in solution paths that are not piecewise linear and for this reason several algorithms have been proposed in the literature. Park and Hastie (2007)

Address for correspondence: Ernst C. Wit, Department of Mathematics, University of Groningen, Bernoulliborg Nijenborgh 9, Groningen 9747 AG, The Netherlands.
E-mail: e.c.wit@rug.nl

 1369–7412/13/75471

proposed an L_1-regularization path following algorithm, and co-ordinate descent methods were proposed in Wu and Lange (2008) and Friedman *et al.* (2010) to improve the computational speed for the path following algorithm. Goeman (2009) proposed a gradient ascent algorithm. Meier *et al.* (2009) proposed a new penalty function for high dimensional generalized additive models, which penalizes both non-sparsity and roughness.

Efron *et al.* (2004) introduced a new method to select important variables in a linear regression model called the *least angle regression* (LARS) algorithm. In the LARS method a multivariate solution path is defined by using the geometrical theory of the linear regression model. Although the LARS algorithm neither optimizes nor penalizes the likelihood, there are close links between it and the lasso. The solution path of the LARS algorithm coincides, with minor adjustments, with the solution path of the lasso obtained by varying the L_1-penalty parameter. Recently, James *et al.* (2009) showed that a LARS-type algorithm, called DASSO, can be also used to produce the entire coefficient path for the Dantzig selector. These relationships suggest that the geometrical structure of a regression model can be used to study its sparse structure in a more direct way. Based on this idea, in this paper we propose a method for GLMs that defines sparsity more directly than through an L_1-penalty on the likelihood. Our approach is theoretically founded on the differential geometrical structure of a GLM and allows us to extend in a natural way the notion of the equiangularity condition that was originally proposed in Efron *et al.* (2004).

The paper is organized as follows. In Section 2 we introduce the differential geometric theory underlying GLMs. Section 3 is devoted to the theoretical aspects of the method proposed; more specifically, in Section 3.1 we use our differential geometric setting to generalize the equiangularity condition for generalized linear regression models, which will be used in Section 3.2 to define the differential geometric LARS (DGLARS) method. The computational aspects are described in Section 3.3 where we propose a predictor corrector algorithm to compute the solution curve. In Section 4 we study two properties of the DGLARS method. In Section 4.1 we use covariance penalty theory to define the notion of generalized degrees of freedom of the method proposed. Using the theory of local co-ordinate systems, we propose an estimator of the generalized degrees of freedom, which we compare with the often used estimator of model complexity, i.e. the cardinality of the active set. In Section 4.2 we study the relationship between DGLARS and L_1-penalized GLMs. In Section 5, we use several simulation studies and a real data set to study and compare the behaviour of the proposed method with some of the most important sparse GLM algorithms. Finally in Section 6 we draw some conclusions.

2. The differential geometry of the generalized linear model

In this section we introduce the GLM (McCullagh and Nelder, 1989) from a differential geometric point of view. In our treatment, we rely heavily on Amari (1985), Kass and Vos (1997) and Amari and Nagaoka (2000). A differential geometric approach was also used in Wei (1998) to study non-linear models based on the exponential family. Essential aspects of differential and information geometry have been included to make the paper self-contained.

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ be a random vector with independent components. In what follows we shall assume that the ith element of $\mathbf{Y}$, Y_i, is a random variable with probability density function belonging to the exponential family

$$p_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}, \qquad y_i \in \mathcal{Y}_i \subseteq \mathbb{R}, \tag{1}$$

where the canonical parameter $\theta_i \in \Theta_i \subseteq \mathbb{R}$, the dispersion parameter $\phi \in \Phi \subseteq \mathbb{R}^+$ and $a(\cdot)$, $b(\cdot)$

and $c(\cdot,\cdot)$ are specific given functions. Under family (1), the joint probability density function of the random vector $\mathbf{Y}$ can be written as

$$p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta},\phi)=\prod_{i=1}^{n} p_{Y_i}(y_i;\theta_i,\phi),$$

where the canonical parameter $\boldsymbol{\theta}$ varies in the subset $\otimes_{i=1}^{n}\Theta_i=\Theta\subseteq\mathbb{R}^n$. The mean value of $\mathbf{Y}$ is denoted by $\boldsymbol{\mu}=(\mu(\theta_1),\ldots,\mu(\theta_n))'$, where $\mu(\theta_i)=\partial b(\theta_i)/\partial\theta$ is called mean value mapping, and the variance of $\mathbf{Y}$ is equal to $\mathrm{var}(\mathbf{Y})=a(\phi)\,\mathbf{V}(\boldsymbol{\mu})$, where $\mathbf{V}(\boldsymbol{\mu})$ is an $n\times n$ diagonal matrix with elements $V(\mu_i)=\partial^2 b(\theta_i)/\partial\theta_i^2$. $V(\cdot)$ is called the variance function. Since $\mu(\cdot)$ is a one-to-one function from $\mathrm{int}(\Theta)$ onto $\tilde{\mathcal{S}}=\mu\{\mathrm{int}(\Theta)\}$, $p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta},\phi)$ may be parameterized by $(\boldsymbol{\mu};\phi)$. Without loss of generality we can assume that $\phi=1$ (Kass and Vos,1997). Assuming that Θ is open, the set

$$\mathcal{S}=\{p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\mu}):\boldsymbol{\mu}\in\tilde{\mathcal{S}}\} \tag{2}$$

is a minimal and regular exponential family of order n and can be treated as a differential manifold where the parameter vector $\boldsymbol{\mu}$ plays the role of a co-ordinate system (Amari, 1985). The notion of differential manifold is necessary for extending the methods of differential calculus to a space that is more general than $\mathbb{R}^n$. For a rigorous definition of a differential manifold the reader is referred to Spivak (1979) and do Carmo (1992). It is worth noting that the results coming from differential geometry are not related to the chosen co-ordinate system, i.e. the parameterization that is used to specify the probability density function (1). This means that we could work with the differential manifold $\mathcal{S}$ using the parameter vector $\boldsymbol{\theta}$ as co-ordinate system. In this paper we prefer to use definition (2) only because we believe that this makes the generalization of the LARS algorithm clearer.

A GLM is completely specified by the following assumptions:

(a) $\mathbf{y}$ is a random observation drawn from the distribution on $\mathbf{Y}$;
(b) for each random variable Y_i there is a vector of covariates $\mathbf{x}_i=(x_{i1},x_{i2},\ldots,x_{ip})'\in\mathcal{X}\subseteq\mathbb{R}^p$;
(c) $E(Y_i|\mathbf{x}_i)=\mu_i(\boldsymbol{\beta})=G(\mathbf{x}_i'\boldsymbol{\beta})$, where $g=G^{-1}$ is called the *link function*.

In James (2002) an interesting extension of the classical GLM is proposed to handle functional predictors. In the literature this model is known as the generalized functional linear model and was also studied in Müller and Stadtmüller (2005) and Li *et al.* (2010), among others.

To simplify our notation, we denote $\boldsymbol{\mu}(\boldsymbol{\beta})=(G(\mathbf{x}_1'\boldsymbol{\beta}),G(\mathbf{x}_2'\boldsymbol{\beta}),\ldots,G(\mathbf{x}_n'\boldsymbol{\beta}))'$.

Assuming that $\boldsymbol{\beta}\to\boldsymbol{\mu}(\boldsymbol{\beta})$ is an embedding, the set

$$\mathcal{M}=\{p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\mu}(\boldsymbol{\beta}))\in\mathcal{S}:\boldsymbol{\beta}\in\mathbb{R}^p\}$$

is a p-dimensional submanifold of $\mathcal{S}$. To obtain a natural generalization of the equiangularity condition that was proposed by Efron *et al.* (2004), it is necessary to introduce two fundamental notions on which Riemannian geometry is based: the notions of a tangent space and a Riemannian metric. To complete the differential geometric setting for the GLM, we shall assume that the usual regularity conditions hold (Amari (1985), page 16). Throughout this paper we use the convention that the indices i, j and k correspond to the quantities that are related to $\boldsymbol{\mu}\in\tilde{\mathcal{S}}$ whereas the indices l, m and q correspond to the quantities that are related to the coefficients $\boldsymbol{\beta}\in\mathbb{R}^p$ of our regression model.

Consider a double-differentiable curve, say $\boldsymbol{\mu}:\Gamma\to\tilde{\mathcal{S}}$, where Γ is the real interval $(-\delta,\delta)$ with $\delta>0$. The tangent vector to the one-parametric family $p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\mu}(\gamma))$ at $\boldsymbol{\mu}=\boldsymbol{\mu}(0)$ is defined as

$$v(\mathbf{Y})=\frac{\mathrm{d}l(\boldsymbol{\mu}(\gamma);\mathbf{Y})}{\mathrm{d}\gamma}\Big|_{\gamma=0}=\sum_{i=1}^{n}\mathrm{d}\mu_i(0)\,\partial_i l(\boldsymbol{\mu};\mathbf{Y}), \tag{3}$$

where $\mathrm{d}\mu_i(0) = \mathrm{d}\mu_i(\gamma)/\mathrm{d}\gamma|_{\gamma=0}$ and $\partial_i l(\boldsymbol{\mu}; \mathbf{Y}) = \partial \log\{p_{\mathbf{Y}}(\mathbf{Y}; \boldsymbol{\mu}(\gamma))\}/\partial\mu_i|_{\gamma=0}$. Roughly speaking, the tangent space of $\mathcal{S}$ at the point $p_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu})$, denoted by $T_{p(\boldsymbol{\mu})}\mathcal{S}$, is the set of all possible tangent vectors at $\boldsymbol{\mu} = \boldsymbol{\mu}(0)$. Formally, $T_{p(\boldsymbol{\mu})}\mathcal{S}$ is the vector space that is spanned by the n score functions $\partial_i l(\boldsymbol{\mu}; \mathbf{Y})$:

$$T_{p(\boldsymbol{\mu})}\mathcal{S} = \mathrm{span}\{\partial_1 l(\boldsymbol{\mu}; \mathbf{Y}), \partial_2 l(\boldsymbol{\mu}; \mathbf{Y}), \ldots, \partial_n l(\boldsymbol{\mu}; \mathbf{Y})\}. \tag{4}$$

Under the regularity conditions cited above, $T_{p(\boldsymbol{\mu})}\mathcal{S}$ is a subspace of squared integrable random variables, in which elements $v(\mathbf{Y})$ satisfy the property $E_{\boldsymbol{\mu}}\{v(\mathbf{Y})\} = 0$, where the expected value is computed with respect to $p_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu})$. As an application of the chain rule, it is easy to see that the definition of a tangent space does not depend on the chosen parameterization; in other words the tangent space can be defined as the vector space that is spanned by the n score functions $\partial_i^* l(\boldsymbol{\theta}; \mathbf{Y}) = \partial \log\{p_{\mathbf{Y}}(\mathbf{Y}; \boldsymbol{\theta}(\gamma))\}/\partial\theta_i|_{\gamma=0}$ where $\boldsymbol{\theta}(\gamma) = \theta(\boldsymbol{\mu}(\gamma))$. Using the terminology that was introduced in Vos (1991), $\partial_i l(\boldsymbol{\mu}; \mathbf{Y})$ are the natural bases of the tangent space when we choose $\boldsymbol{\mu}$ as co-ordinate system, whereas $\partial_i^* l(\boldsymbol{\theta}; \mathbf{Y})$ are the natural bases when $\boldsymbol{\theta}$ is used as the co-ordinate system.

Similarly, consider a double-differentiable curve $\boldsymbol{\beta}: \Gamma' \to \mathbb{R}^p$, with $\Gamma' = (-\delta', \delta')$ and $\delta' > 0$. The tangent vector to the one-parametric family $p_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu}(\boldsymbol{\beta}(\gamma)))$ at the point $\boldsymbol{\beta} = \boldsymbol{\beta}(0)$ is defined as

$$w(\mathbf{Y}) = \sum_{m=1}^{p} \mathrm{d}\beta_m(0)\, \partial_m l(\boldsymbol{\beta}; \mathbf{Y}),$$

where $\mathrm{d}\beta_m(0) = \mathrm{d}\beta_m(\gamma)/\mathrm{d}\gamma|_{\gamma=0}$ and $\partial_m l(\boldsymbol{\beta}; \mathbf{Y}) = \partial \log\{p_{\mathbf{Y}}(\mathbf{Y}; \boldsymbol{\mu}(\boldsymbol{\beta}(\gamma)))\}/\partial\beta_m|_{\gamma=0}$. Then, the tangent space of $\mathcal{M}$ at the point $p_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu}(\boldsymbol{\beta}))$ is

$$T_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}}\mathcal{M} = \mathrm{span}\{\partial_1 l(\boldsymbol{\beta}; \mathbf{Y}), \partial_2 l(\boldsymbol{\beta}; \mathbf{Y}), \ldots, \partial_p l(\boldsymbol{\beta}; \mathbf{Y})\}.$$

The definition of the inner product on each tangent space allows us to generalize the notion of angle between two curves, say $\boldsymbol{\mu}_1(\gamma)$ and $\boldsymbol{\mu}_2(\gamma)$, intersecting at $\boldsymbol{\mu}_1(0) = \boldsymbol{\mu}_2(0) = \boldsymbol{\mu}$, with tangent vectors belonging to $T_{p(\boldsymbol{\mu})}\mathcal{S}$, denoted by

$$v_1(\mathbf{Y}) = \sum_{i=1}^{n} \mathrm{d}\mu_{1,i}(0)\, \partial_i l(\boldsymbol{\mu}; \mathbf{Y})$$

and

$$v_2(\mathbf{Y}) = \sum_{i=1}^{n} \mathrm{d}\mu_{2,i}(0)\, \partial_i l(\boldsymbol{\mu}; \mathbf{Y})$$

respectively. When working with a parametric family of distributions, the inner product can be defined in a natural way (Rao, 1945), i.e.

$$\langle v_1(\mathbf{Y}), v_2(\mathbf{Y}) \rangle_{p(\boldsymbol{\mu})} = E_{\boldsymbol{\mu}}\{v_1(\mathbf{Y})\, v_2(\mathbf{Y})\} = \mathrm{d}\boldsymbol{\mu}_1(0)'\, I(\boldsymbol{\mu})\, \mathrm{d}\boldsymbol{\mu}_2(0),$$

where $I(\boldsymbol{\mu})$ is the Fisher information matrix for the mean parameter at point $\boldsymbol{\mu}$. In other words, the Fisher information defines a Riemannian metric by associating with each point of $\mathcal{S}$ an inner product on the tangent space. This Riemannian metric is also called the *information metric* (Burbea and Rao, 1982). Since $T_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}}\mathcal{M}$ is a linear subspace of $T_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}}\mathcal{S}$, the Fisher information also defines an inner product on $T_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}}\mathcal{M}$. Therefore, we can define the inner product between a tangent vector $w(\mathbf{Y})$ of $T_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}}\mathcal{M}$ and a tangent vector $v(\mathbf{Y})$ of $T_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}}\mathcal{S}$, namely

$$\langle w(\mathbf{Y}), v(\mathbf{Y}) \rangle_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}} = E_{\boldsymbol{\mu}(\boldsymbol{\beta})}\{w(\mathbf{Y})\, v(\mathbf{Y})\} = \mathrm{d}\boldsymbol{\beta}(0)' \frac{\partial \boldsymbol{\mu}(\boldsymbol{\beta})'}{\partial \boldsymbol{\beta}} I\{\boldsymbol{\mu}(\boldsymbol{\beta})\}\, \mathrm{d}\boldsymbol{\mu}(0),$$

where $\partial\boldsymbol{\mu}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is the Jacobian matrix of the vector function $\boldsymbol{\mu}(\boldsymbol{\beta})$.

Each Riemannian metric defines the notion of a geodesic, i.e. the generalization of a straight line in a differential geometric framework. Roughly speaking, a geodesic can be defined as the shortest path between two given points on a differential manifold. A geodesic is defined as the solution of a system of differential equations, the Euler–Lagrange equations, obtained from defining a connection on a differentiable manifold. In statistical theory a one-parametric family of connections plays a fundamental role, the so-called α-connections, denoted by ∇^{α}, that generalize the classical notion of a Levi–Civita connection, which is the special case that $\alpha = 0$. In the theory of information geometry, ∇^{0} is also called the *information connection* since it is derived from the Fisher information. In Section 4.2 we shall use the Levi–Civita connection to link the method proposed to the L_1-penalized GLM. What is also important for following this paper is that $\mathcal{S}$ is a dually flat space, namely, it is flat with respect to the 1- and -1-connection. In this paper we shall not discuss the details of this dual geometry. For a complete treatment the reader is referred to Amari and Nagaoka (2000). As shown in Vos (1991), associated with the -1-connection and each point $p_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu})$ there is a diffeomorphism between a neighbourhood of the origin in $T_{p(\boldsymbol{\mu})}\mathcal{S}$ and a neighbourhood of $p_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu})$, called the -1-*exponential map*. The dual nature that exists between ∇^{-1} and ∇^{1} defines the dual of the -1-exponential map, namely the so-called 1-exponential map. Since $\mathcal{S}$ is a dually flat space, the inverses of the two exponential maps are well defined. To complete the geometrical framework that is needed to generalize the LARS algorithm, we consider the inverse of the -1-exponential map, which relates the observed response variable $\mathbf{y}$ to the tangent spaces. Vos (1991) defined what we call the *tangent residual vector*

$$\mathbf{r}(\boldsymbol{\mu}(\boldsymbol{\beta}), \mathbf{y}; \mathbf{Y}) = \sum_{i=1}^{n} \{y_i - \mu_i(\boldsymbol{\beta})\}\, \partial_i l(\boldsymbol{\mu}(\boldsymbol{\beta}); \mathbf{Y}) \tag{5}$$

where $\partial_i l(\boldsymbol{\mu}(\boldsymbol{\beta}); \mathbf{Y}) = \partial l(\boldsymbol{\mu}; \mathbf{Y})/\partial \mu_i|_{\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\beta})}$. We define the tangent residual vector (5) with respect to both the fixed observations $\mathbf{y}$ and the random variable $\mathbf{Y}$, in such a way that it is a random variable with zero expected value and finite variance, and therefore $\mathbf{r}(\boldsymbol{\mu}(\boldsymbol{\beta}), \mathbf{y}; \mathbf{Y}) \in T_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}}\mathcal{S}$. Vos (1991) showed that it is possible to give a differential geometric interpretation of the maximum likelihood estimator by using the tangent residual vector and the tangent space $T_{p\{\boldsymbol{\mu}(\hat{\boldsymbol{\beta}})\}}\mathcal{M}$, namely $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$ when the tangent residual vector is orthogonal to the tangent space $T_{p\{\boldsymbol{\mu}(\hat{\boldsymbol{\beta}})\}}\mathcal{M}$. It is worth noting that this statement is well defined even if $\mathbf{y}$ is not an element of the mean value parameter space $\tilde{\mathcal{S}}$. In other words, the differential geometric description of the maximum likelihood estimator can be used even if the Kullback–Leibler divergence is not defined (Vos, 1991).

3. Differential geometric least angle regression

A GLM relates a linear combination of covariates via a link function to the distribution of the observations. If it is not known which covariates are actually predictive for the outcome, various procedures have been proposed to zoom in on the most relevant features. A large group of stepwise procedures were the first attempt to 'select' variables (Hocking, 1976). Principal component regression (Jolliffe, 1982) was a recognition of the fact that similar information could be present in several variables. The lasso (Tibshirani, 1996) heralded the era of path algorithms, which often indirectly select variables owing to a *pleasant coincidence* between the geometry of the model and the choice of penalty. Least angle regression (Efron *et al.*, 2004) was originally intended as a computational tool. In this paper, however, we shall present this algorithm as a principled method for directly connecting the geometry of the model to the sparsity of the feature space. In other words, least angle regression is not only 'an important contribution to

Table 1. Overview of the DGLARS method to compute the solution curve

Step	*Algorithm*
1	Start with the intercept-only model
2	Repeat
3	Increase the parameters of the active variables keeping the angles between their scores and residual tangent vector the same
4	If the angle of a not-included variable is the same as the ones currently in the model include that variable in the active set
5	Until a stopping rule is met

statistical computing' (Madigan and Ridgeway, 2004) but also a new method in its own right: it can be generalized to any model and its success does not depend on the arbitrary match of the constraint and the objective function.

The original LARS algorithm defines a solution path of a linear regression model by sequentially adding variables to the solution. Starting with only the intercept, the LARS algorithm finds the covariate that is most correlated with the response variable and proceeds in this 'direction' by changing its associated linear parameter. The algorithm takes the largest step possible in the direction of this covariate until some other covariate has as much correlation with the current residual as the current covariate. At that point the LARS algorithm proceeds in an equiangular direction between the two covariates until a new covariate earns its way into the *equally most correlated set*. Then it proceeds in the direction in which the residual makes an equal angle with the three covariates, and so on. For an extensive review of this method, the reader is referred to Hesterberg *et al.* (2008). In this section we generalize these notions for GLMs by using differential geometry. Table 1 gives an overview of the method.

3.1. Equiangularity in a generalized linear model

In the linear regression model, the notion of the angle between the covariates and the residual is independent from the form of the model space simply because the model is defined as the collection of linear combinations of the covariates. The linearity of the models results in the piecewise linearity of the LARS solution paths. For a GLM, the effect of any covariate on the residual is moderated by the link function and the parameterization. In this section we describe how the geometrical setting that was introduced in Section 2 can be used to define a genuine generalization of the LARS algorithm for GLMs. In what follows we shall assume that all models include an intercept.

Let $\hat{\beta}_{a_0}$ be the maximum likelihood estimate of the intercept β_{a_0} within the intercept-only log-likelihood $l(\boldsymbol{\mu}(\beta_{a_0}); \mathbf{y})$, which is used as the starting point of the proposed generalization. In our approach the use of the maximum likelihood estimator is limited to the starting point. As noted above, the tangent residual vector $\mathbf{r}(\boldsymbol{\mu}(\hat{\beta}_{a_0}), \mathbf{y}; \mathbf{Y})$ is orthogonal to the basis $\partial_{a_0} l(\hat{\beta}_{a_0}; \mathbf{Y})$ of the tangent space $T_{p\{\boldsymbol{\mu}(\hat{\beta}_{a_0})\}}\mathcal{M}$. The tangent residual vector can be used to rank the covariates locally by using the notion of angle defined on the tangent space. As shown in Fig. 1(a), the method proposed finds that covariate, say $\mathbf{x}_{a_1}$, whose basis vector $\partial_{a_1} l(\hat{\beta}_{a_0}; \mathbf{Y})$ has the smallest angle with the tangent residual vector.

The method then includes the covariate $\mathbf{x}_{a_1}$ in the active set $\mathcal{A}(\gamma^{(1)}) = \{a_0, a_1\}$. The solution curve $\boldsymbol{\beta}(\gamma) = (\beta_{a_0}(\gamma), \beta_{a_1}(\gamma))'$ is chosen in such a way that it satisfies the condition that the tangent residual vector is always orthogonal to the basis $\partial_{a_0} l(\boldsymbol{\beta}(\gamma); \mathbf{Y})$. The direction of the curve $\boldsymbol{\beta}(\gamma)$ is defined by the projection of the tangent residual vector on the basis vector $\partial_{a_1} l(\boldsymbol{\beta}(\gamma); \mathbf{Y})$.

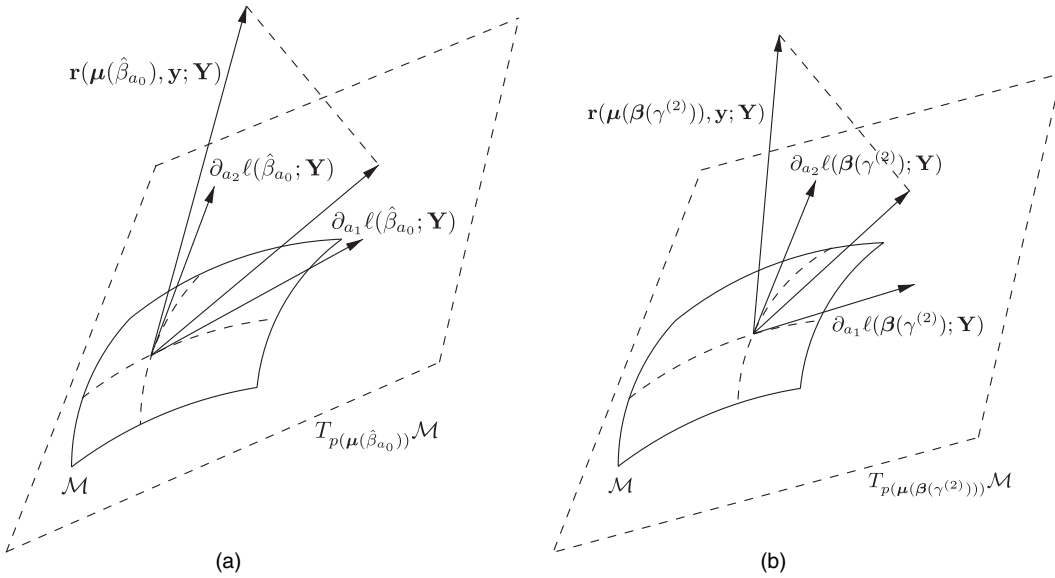

Fig. 1. Differential geometrical description of the LARS algorithm for a GLM with two covariates: (a) the first covariate x_{a_1} is found and included in the active set; (b) the generalized equiangularity condition (10) is satisfied for variables x_{a_1} and x_{a_2}

The curve $\boldsymbol{\beta}(\gamma)$ continues as defined above until $\gamma^{(2)}$, for which there is a covariate, say $\mathbf{x}_{a_2}$, that satisfies the equiangularity condition on the tangent space $T_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma^{(2)}))\}}\mathcal{M}$, in other words

$$\rho_{a_1}\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma^{(2)}))\} = \rho_{a_2}\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma^{(2)}))\},$$

where $\rho_m\{\boldsymbol{\mu}(\boldsymbol{\beta})\}$ is the angle between the tangent residual vector and the basis vector $\partial_m l(\boldsymbol{\beta}(\gamma); \mathbf{Y})$. At this point $\mathbf{x}_{a_2}$ is included in the active set $\mathcal{A}(\gamma^{(2)})$ and a new curve $\boldsymbol{\beta}(\gamma) = (\beta_{a_0}(\gamma), \beta_{a_1}(\gamma), \beta_{a_2}(\gamma))'$ is defined, such that the tangent residual vector is always orthogonal to the basis vector $\partial_{a_0} l(\boldsymbol{\beta}(\gamma); \mathbf{Y})$ with direction defined by the tangent vector that bisects the angle between the basis vectors $\partial_{a_1} l(\boldsymbol{\beta}(\gamma); \mathbf{Y})$ and $\partial_{a_2} l(\boldsymbol{\beta}(\gamma); \mathbf{Y})$, as shown in Fig. 1(b).

We note that, in principle, we treat the intercept differently from the other covariates. Unless there are some special reasons to do otherwise, the intercept will always be included. Therefore, we do not 'penalize' the intercept, in the sense that the tangent residual vector is constrained to be always orthogonal to the basis vector $\partial_{a_0} l(\boldsymbol{\beta}(\gamma); \mathbf{Y})$. This means that the tangent residual vector contains only information on the covariates. Although the proposed generalization is based on the idea of using $\hat{\beta}_{a_0}$ as the starting point, when $\boldsymbol{\mu}(0) \in \tilde{\mathcal{S}}$, it can be modified to deal with models without the intercept term. In this case $\mathbf{r}(\boldsymbol{\mu}(0), \mathbf{y}; \mathbf{Y})$ is used to rank the covariates locally. This modification can be used for several important models such as the logistic regression model and the Poisson regression model, in both cases with and without an intercept term.

3.2. Formal description of differential geometric least angle regression

The derivative of the log-likelihood $l(\boldsymbol{\beta}(\gamma); \mathbf{y})$ with respect to the mth covariate parameter can be written as the inner product between the current tangent residual vector $\mathbf{r}(\boldsymbol{\beta}(\gamma), \mathbf{y}; \mathbf{Y})$ and the mth base of the tangent space of $\mathcal{M}$,

$$\partial_m l(\boldsymbol{\beta}(\gamma); \mathbf{y}) = \langle \partial_m l(\boldsymbol{\beta}(\gamma); \mathbf{Y}); \mathbf{r}(\boldsymbol{\beta}(\gamma), \mathbf{y}; \mathbf{Y}) \rangle_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}. \tag{6}$$

Using the law of cosines, this expression is equivalent to

$$\begin{aligned}\partial_m l(\boldsymbol{\beta}(\gamma);\mathbf{y}) &= \cos[\rho_m\{\boldsymbol{\beta}(\gamma)\}]\,\|\mathbf{r}(\boldsymbol{\beta}(\gamma),\mathbf{y};\mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}\,\|\partial_m l(\boldsymbol{\beta}(\gamma);\mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}} \\ &= \cos[\rho_m\{\boldsymbol{\beta}(\gamma)\}]\,\|\mathbf{r}(\boldsymbol{\beta}(\gamma),\mathbf{y};\mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}\, i_m^{1/2}\{\boldsymbol{\beta}(\gamma)\}, \end{aligned} \tag{7}$$

where $\rho_m\{\boldsymbol{\beta}(\gamma)\}$ is the angle between the tangent residual vector $\mathbf{r}(\boldsymbol{\beta}(\gamma),\mathbf{y};\mathbf{Y})$ and the mth basis function $\partial_m l(\boldsymbol{\beta}(\gamma);\mathbf{Y})$, $\|\cdot\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}$ is the norm defined on $T_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}}\mathcal{M}$ and $i_m\{\boldsymbol{\beta}(\gamma)\}$ is the Fisher information for $\beta_m(\gamma)$. Importantly, equation (7) shows that the gradient of the log-likelihood function does not generalize the equiangularity condition that was proposed in Efron *et al.* (2004) to define the LARS algorithm, since the latter does not consider the variation related to the square root of the Fisher information $i_m^{1/2}\{\boldsymbol{\beta}(\gamma)\}$, which in the case of a GLM is typically not constant. Using equation (7), the angle $\rho_m\{\boldsymbol{\beta}(\gamma)\}$ can be written as

$$\rho_m\{\boldsymbol{\beta}(\gamma)\} = \cos^{-1}\left[\frac{\partial_m l(\boldsymbol{\beta}(\gamma);\mathbf{y})}{\|\mathbf{r}(\boldsymbol{\beta}(\gamma),\mathbf{y};\mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}\, i_m^{1/2}\{\boldsymbol{\beta}(\gamma)\}}\right].$$

We can define the equiangularity condition directly on $\rho_m\{\boldsymbol{\beta}(\gamma)\}$ as in the case of LARS, but it is easier and more intuitive to define the same condition on a transformation of the same quantity. Let $r_m^u(\gamma)$ be the signed *Rao score test statistic*, where

$$r_m^u(\gamma) = i_m^{-1/2}\{\boldsymbol{\beta}(\gamma)\}\,\partial_m l(\boldsymbol{\beta}(\gamma);\mathbf{y}) \tag{8}$$

$$= \cos[\rho_m\{\boldsymbol{\beta}(\gamma)\}]\,\|\mathbf{r}(\boldsymbol{\beta}(\gamma),\mathbf{y};\mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}. \tag{9}$$

Note that the inverse cosine is a strictly increasing function on its restricted domain and that $\|\mathbf{r}(\boldsymbol{\beta}(\gamma),\mathbf{y};\mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}$ does not depend on m. Therefore, the signed Rao score test statistic contains the same information as the angle $\rho_m\{\boldsymbol{\beta}(\gamma)\}$. As a result, for GLMs we can define DGLARS with respect to the Rao score test statistics, rather than the angles.

Furthermore, we note that the Rao score test statistic as defined by equation (9) breaks down into a variable selection part, $\cos[\rho_m\{\boldsymbol{\beta}(\gamma)\}]$, and a measure of global fit of the model, $\|\mathbf{r}(\boldsymbol{\beta}(\gamma),\mathbf{y};\mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}$. In contrast, in the form (8), the Rao score test statistic stresses its invariance to any one-to-one reparameterization of the form $\zeta_m = \zeta_m(\beta_m)$. In Efron *et al.* (2004) this aspect is not treated, because they assumed that for all m the information $i_m(\boldsymbol{\beta})$ is equal to 1. In this way they could drop $i_m^{-1/2}\{\boldsymbol{\beta}(\gamma)\}$ and focus only on the derivative of the log-likelihood function, i.e. the covariance between x_m and the tangent residual vector.

The solution curve, which is denoted by $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma) \subset \mathbb{R}^{k+1}$, with $\gamma \in [0, \gamma^{(1)}]$, whereby

$$0 \leqslant \gamma^{(p)} \leqslant \ldots \leqslant \gamma^{(2)} \leqslant \gamma^{(1)},$$

is defined in the following way: for any $\gamma \in (\gamma^{(k+1)}; \gamma^{(k)}]$, $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)$ is chosen in such a way that

$$\left.\begin{array}{ll} & \mathcal{A}(\gamma) = \{a_1, a_2, \ldots, a_k\}, \\ |r_{a_i}^u(\gamma)| = |r_{a_j}^u(\gamma)|, & \forall a_i, a_j \in \mathcal{A}(\gamma), \\ |r_{a_h^c}^u(\gamma)| < |r_{a_i}^u(\gamma)|, & \forall a_h^c \in \mathcal{A}^c(\gamma) \text{ and } \forall a_i \in \mathcal{A}(\gamma). \end{array}\right\} \tag{10}$$

In what follows we shall call expression (10) the *generalized equiangularity condition*. When $\gamma = \gamma^{(j)}$, with $j = 2, \ldots, p$, the following condition is satisfied:

$$\exists a_h^c \in \mathcal{A}^c(\gamma) : |r_{a_h^c}^u(\gamma^{(j)})| = |r_{a_i}^u(\gamma^{(j)})|, \qquad \forall a_i \in \mathcal{A}(\gamma); \tag{11}$$

in this case a new covariate is included in the active set.

3.3. Predictor–corrector algorithm

To compute the solution curve we use the predictor–corrector algorithm (Allgower and Georg, 2003). The basic idea underlying the predictor–corrector algorithm is to trace a curve implicitly defined by a system of non-linear equations. The curve is obtained by generating a sequence of points satisfying a chosen tolerance criterion. A predictor–corrector algorithm was also used in Park and Hastie (2007) to compute the path of the coefficients of a GLM with L_1-penalty function.

Let us suppose that k covariates are included in the active set, $\mathcal{A}(\gamma) = \{a_1, a_2, \ldots, a_k\}$. Using the generalized equiangularity condition (10), the solution curve satisfies the relationship

$$|r^u_{a_1}(\gamma)| = |r^u_{a_2}(\gamma)| = \ldots = |r^u_{a_k}(\gamma)|, \tag{12}$$

for any $\gamma \in (\gamma^{(k+1)}; \gamma^{(k)}]$. Let $\mathbf{v}_k = \{v_{a_1}, v_{a_2}, \ldots, v_{a_k}\}$ be the vector such that $v_{a_j} = \text{sgn}\{r^u_{a_j}(\gamma^{(k)})\}$; the solution curve $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)$ is implicitly defined by the following system of $k+1$ non-linear equations:

$$\left.\begin{array}{cc} \partial_{a_0} l(\boldsymbol{\beta}(\gamma); \mathbf{y}) = 0, & \\ r^u_{a_1}(\gamma) = v_{a_1}\gamma, & \\ \vdots \qquad\qquad \vdots & \\ r^u_{a_k}(\gamma) = v_{a_k}\gamma. & \end{array}\right\} \tag{13}$$

When $\gamma = 0$ we obtain the maximum likelihood estimates of the subset of the parameter vector $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$, of the covariates in the active set. The point $\hat{\boldsymbol{\beta}}(\gamma^{(k+1)})$ lies on the solution curve joining $\hat{\boldsymbol{\beta}}(\gamma^{(k)})$ with $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$. To simplify our notation, we define $\tilde{\boldsymbol{\varphi}}_{\mathcal{A}}(\gamma) = \boldsymbol{\varphi}_{\mathcal{A}}(\gamma) - \mathbf{v}_{\mathcal{A}}\gamma$, where $\boldsymbol{\varphi}_{\mathcal{A}}(\gamma) = (\partial_{a_0} l(\boldsymbol{\beta}(\gamma); \mathbf{y}), r^u_{a_1}(\gamma), \ldots, r^u_{a_k}(\gamma))'$ and $\mathbf{v}_{\mathcal{A}} = (0, \mathbf{v}_k)'$. If the model is with no intercept and satisfies the condition that was seen at the end of Section 3.1, then $\boldsymbol{\varphi}_{\mathcal{A}}(\gamma) = (r^u_{a_1}(\gamma), \ldots, r^u_{a_k}(\gamma))'$ and $\mathbf{v}_{\mathcal{A}} = \mathbf{v}_k$. By differentiating $\tilde{\boldsymbol{\varphi}}_{\mathcal{A}}(\gamma)$ with respect to γ, we obtain

$$\frac{\mathrm{d}\tilde{\boldsymbol{\varphi}}_{\mathcal{A}}(\gamma)}{\mathrm{d}\gamma} = \frac{\partial \boldsymbol{\varphi}_{\mathcal{A}}(\gamma)}{\partial \hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)} \frac{\mathrm{d}\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)}{\mathrm{d}\gamma} - \mathbf{v}_{\mathcal{A}} = \mathbf{0}, \tag{14}$$

where $\partial \boldsymbol{\varphi}_{\mathcal{A}}(\gamma)/\partial \hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)$ is the Jacobian matrix of the vector function $\boldsymbol{\varphi}_{\mathcal{A}}(\gamma)$ evaluated at the point $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)$. Using expression (14), we can locally approximate the solution curve by the expression

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma - \Delta\gamma) \approx \tilde{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma - \Delta\gamma) = \hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma) - \frac{\partial \boldsymbol{\varphi}_{\mathcal{A}}(\gamma)}{\partial \hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)}^{-1} \mathbf{v}_{\mathcal{A}} \Delta\gamma, \tag{15}$$

where $\Delta\gamma \in [0; \gamma - \gamma^{(k+1)}]$. We use expression (15) for the predictor step of the proposed algorithm. An efficient implementation of the predictor–corrector method requires a suitable method to compute the step size $\Delta\gamma$. Several methods have been proposed in the literature to solve this problem. For example, we can consider a fixed value of $\Delta\gamma$ or we can relate the step size with a fixed variation in the arc length parameterization of the solution curve (see chapter 6 in Allgower and Georg (2003) for further details). In this paper, we use the method that was proposed in Park and Hastie (2007), namely we consider the step size that changes the active set. Using expression (11), we have a change in the active set when

$$\exists a^{\mathrm{c}}_h \in \mathcal{A}^{\mathrm{c}}(\gamma) : |r^u_{a^{\mathrm{c}}_h}(\gamma - \Delta\gamma)| = |r^u_{a_i}(\gamma - \Delta\gamma)|, \qquad \forall a_i \in \mathcal{A}(\gamma). \tag{16}$$

Expanding $r^u_{a^{\mathrm{c}}_h}(\gamma)$ in a Taylor series around γ, we consider the expression

$$|r^u_{a^c_h}(\gamma - \Delta\gamma)| \approx \left| r^u_{a^c_h}(\gamma) - \frac{\mathrm{d}r^u_{a^c_h}(\gamma)}{\mathrm{d}\gamma}\Delta\gamma \right|.$$

Then, observing that the solution curve satisfies system (13), it is easy to see that the following identity holds:

$$|r^u_{a_i}(\gamma - \Delta\gamma)| = (\gamma - \Delta\gamma), \qquad \Delta\gamma \in [0; \gamma].$$

By combining these two results, condition (16) can be rewritten in the following way:

$$\exists a^c_h \in \mathcal{A}^c(\gamma) : \left| r^u_{a^c_h}(\gamma) - \frac{\mathrm{d}r^u_{a^c_h}(\gamma)}{\mathrm{d}\gamma}\Delta\gamma \right| \approx \gamma - \Delta\gamma, \qquad \forall a_i \in \mathcal{A}(\gamma) \text{ and } \Delta\gamma \in [0; \gamma]$$

and then we have

$$\Delta\gamma^{\mathrm{opt}} = \min^{+}_{a^c_h \in \mathcal{A}^c(\gamma)} \left\{ \frac{\gamma - r^u_{a^c_h}(\gamma)}{1 - \mathrm{d}r^u_{a^c_h}(\gamma)/\mathrm{d}\gamma}; \frac{\gamma + r^u_{a^c_h}(\gamma)}{1 + \mathrm{d}r^u_{a^c_h}(\gamma)/\mathrm{d}\gamma} \right\}. \tag{17}$$

Expression (17) generalizes the step size that was proposed in Efron *et al.* (2004).

Since the optimal step size is based on a local approximation, we also include an exclusion step for removing incorrectly included variables in the model. When an incorrect variable is included in the model after the corrector step, we have that there is a non-active variable such that the absolute value of the corresponding Rao score test statistic is greater than γ. Checking this is trivial. To overcome this drawback, the 'optimal' step size from the previous step is reduced by using a small positive constant ε and the inclusion step is redone until the correct variable is joined to the model. A possible choice for ε could be a half of $\Delta\gamma^{\mathrm{opt}}$. In Table 2 we report the pseudocode of the algorithm that was proposed in this section for a model with the intercept.

Table 2. Pseudocode of the developed algorithm to compute the solution curve defined by the DGLARS method for a model with the intercept†

Step	*Algorithm*
1	Compute $\hat{\beta}_{a_0}$
2	$\mathcal{A} = \arg\max_{a^c_j \in \mathcal{A}^c} \lvert r^u_{a^c_j}(\hat{\beta}_{a_0})\rvert$ and $\gamma = \lvert r^u_{a_1}(\hat{\beta}_{a_0})\rvert$
3	Repeat
4	Use equation (17) to compute $\Delta\gamma^{\mathrm{opt}}$ and set $\gamma = \gamma - \Delta\gamma^{\mathrm{opt}}$
5	Use equation (15) to compute $\tilde{\beta}_{\mathcal{A}}(\gamma)$ (*predictor step*)
6	Use $\tilde{\beta}_{\mathcal{A}}(\gamma)$ as the starting point to solve system (13) (*corrector step*)
7	For all $a^c_h \in \mathcal{A}^c$ compute $r^u_{a^c_h}(\gamma)$
8	If $\exists a^c_h \in \mathcal{A}^c$ such that $\lvert r^u_{a^c_h}(\gamma)\rvert > \gamma$ then
9	$\gamma = \gamma + \varepsilon$, with ε a small positive constant, and go to step 5
10	If condition (11) is satisfied update $\mathcal{A}$
11	Until stopping rule is met

†The computational complexity of the algorithm is roughly $O(np^{2.376}\min\{n, p\})$, where n is the number of observations and p the number of variables.

From an inspection of the algorithm, it is clear that computationally the most expensive steps are solving the system of equations in expression (13) and taking the inverse in equation (14). These steps have complexity $O(|\mathcal{A}|^3)$ in a naive implementation, but which can be improved to $O(|\mathcal{A}|^{2.376})$ according to the Coppersmith–Winograd algorithm. Furthermore, iteration across the active set variables results in a total computational complexity of $O(np^{2.376}\min\{n, p\})$, where p is the number of variables and n the number of observations. This compares with a complexity of $O(np\min\{n, p\})$ for the original LARS algorithm.

4. Properties of the differential geometric least angle regression method

In this section we focus on properties of the DGLARS method. First, we look into the issue of complexity of the models that are selected via DGLARS. We derive an estimator of the model complexity with good performance with respect to standard estimators that have been widely reported in the literature. In Section 4.2 we show that there is a relationship between DGLARS and L_1-penalized inference, but only when the space $\mathcal{M}$ is flat with respect to the 0-connection, i.e. the Levi–Civita connection.

4.1. Model complexity and degrees of freedom

The behaviour of the method proposed is closely related to the way of selecting the optimal value of the tuning parameter γ. For the lasso estimator, Zou *et al.* (2007) developed an adaptive model selection criterion to select the regularization parameter on the basis of a rigorous definition of degrees of freedom. Within the classical theory of linear regression models, it is well known that the degrees of freedom are equal to the number of covariates but for non-linear modelling procedures this equivalence is not satisfied.

To define an adaptive model selection criterion, it is of both theoretical and practical relevance to derive the degrees of freedom of the DGLARS method. In this section we propose a differential geometric approach based on *covariance penalty theory*, which was developed by Efron (2004), which gives us a rigorous definition of degrees of freedom for a general modelling procedure.

Covariance penalty theory is theoretically founded on Bregman divergence (Bregman, 1967) to evaluate the prediction behaviour of a general modelling procedure, i.e. a mapping from $\mathcal{Y}$ to $\tilde{\mathcal{S}}$, say φ, that produces an n-dimensional vector of fitted values, denoted by $\hat{\boldsymbol{\mu}}^{\varphi}(\mathbf{y})$ (Ye, 1998). Let $q(\cdot)$ be a concave real-valued function; the Bregman divergence of y_i to $\hat{\mu}_i^{\varphi}(\mathbf{y})$ generated by the concave function $q(\cdot)$ is defined as

$$Q\{y_i, \hat{\mu}_i^{\varphi}(\mathbf{y})\} = q\{\hat{\mu}_i^{\varphi}(\mathbf{y})\} + \partial q\{\hat{\mu}_i^{\varphi}(\mathbf{y})\}\{y_i - \hat{\mu}_i^{\varphi}(\mathbf{y})\} - q(y_i),$$

where $\partial q\{\hat{\mu}_i^{\varphi}(\mathbf{y})\} = \partial q(y_i)/\partial y_i|_{y_i=\hat{\mu}_i^{\varphi}(\mathbf{y})}$. The total *apparent error* of the general modelling procedure φ is measured by summing all the component errors for each observation:

$$\mathrm{err}(\mathbf{y}) = Q\{\mathbf{y}, \hat{\boldsymbol{\mu}}^{\varphi}(\mathbf{y})\} = \sum_{i=1}^{n} Q\{y_i, \hat{\mu}_i^{\varphi}(\mathbf{y})\}. \tag{18}$$

In our approach we use, as suggested by Efron (1986), $q(y) = 2[b\{\theta(y)\} - y\theta(y)]/a(\phi)$ as generating function. In this way, equation (18) is the well-known residual deviance, which is an optimistic assessment of how the fitted model performs on future data. Following Efron (2004), the *predictive error* of $\hat{\boldsymbol{\mu}}^{\varphi}(\mathbf{y})$ is defined as

$$\mathrm{Err}(\mathbf{y}) = E_{\tilde{\mathbf{Y}}}[Q\{\tilde{\mathbf{Y}}, \hat{\boldsymbol{\mu}}^{\varphi}(\mathbf{y})\}],$$

where the expectation is computed with respect to $\tilde{\mathbf{Y}}$, which is an independent copy of $\mathbf{Y}$ and

where $\mathbf{y}$ is considered fixed. The *optimism theorem* (Efron, 2004), which is reported below, relates the expected values of $\mathrm{Err}(\mathbf{y})$ and $\mathrm{err}(\mathbf{y})$.

Theorem 1 (optimism theorem). Let $q(y) = 2[b\{\theta(y)\} - y\theta(y)]/a(\phi)$ be the generating function of the Bregman divergence; then we have the relationship

$$E_{\mathbf{Y}}\{\mathrm{Err}(\mathbf{Y})\} = E_{\mathbf{Y}}\{\mathrm{err}(\mathbf{Y})\} + 2\Omega_{\varphi} \tag{19}$$

where $\Omega_{\varphi} = a(\phi)^{-1} \sum_{i=1}^{n} \mathrm{cov}\{\hat{\theta}_i^{\varphi}(\mathbf{Y}), Y_i\}$ and $\hat{\theta}_i^{\varphi}(\mathbf{Y}) = \theta\{\hat{\mu}_i^{\varphi}(\mathbf{Y})\}$.

Theorem 1 shows that the residual deviance is a biased estimator of the expected value of the true prediction error, but the bias can be removed by using the term Ω_{φ} in equation (19).

The idea to use Ω_{φ} to define the generalized degrees of freedom of a modelling procedure based on the exponential family is due to Shen *et al.* (2004). The theoretical foundation of this definition relies on the general derivation of the Akaike information criterion for an exponential family. As explained in section 6.5 in Burnham and Anderson (2000), the general derivation of the Akaike information criterion for the exponential family is based on finding an unbiased estimator of the quantity

$$-2T = -2\, E_{\mathbf{Y}}\{l(\hat{\boldsymbol{\mu}}^{\varphi}(\mathbf{Y}); \mathbf{Y})\} + \frac{2}{a(\phi)} \sum_{i=1}^{n} \mathrm{cov}\{\hat{\theta}_i^{\varphi}(\mathbf{Y}), Y_i\}, \tag{20}$$

where $T = E_{\mathbf{Y}}[E_{\tilde{\mathbf{Y}}}\{l(\hat{\boldsymbol{\mu}}^{\varphi}(\mathbf{Y}); \tilde{\mathbf{Y}})\}]$. It is important to observe that the exact result (20) does not depend on the estimator used. When we work with the maximum likelihood estimator, standard results from the Akaike information criterion theory for an exponential family tell us that Ω_{φ} is approximately equal to the number of parameters p, i.e. the penalty term that is used to define the Akaike information criterion. For non-linear models, the cardinality of the active set is not necessarily a good estimator for Ω_{φ} and, instead, we propose the following definition of generalized degrees of freedom of the DGLARS method.

Definition 1. Let γ be a fixed value and let $\hat{\boldsymbol{\theta}}(\mathbf{y}; \gamma) = \hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\mathbf{y}; \gamma))$ be the corresponding DGLARS estimate. The generalized degrees of freedom of the DGLARS method are defined as

$$\mathrm{gdf}(\gamma) = \frac{1}{a(\phi)} \sum_{i=1}^{n} \mathrm{cov}\{\hat{\theta}_i(\mathbf{Y}; \gamma); Y_i\} = \sum_{i=1}^{n} V(\mu_i) \frac{\partial E_{\mathbf{Y}}\{\hat{\theta}_i(\mathbf{Y}; \gamma)\}}{\partial \mu_i}. \tag{21}$$

Identity (21) is obtained by using theorem 1 in Shen *et al.* (2004). As a consequence of the relationship between the maximum likelihood estimator and the term Ω_{φ}, when $\gamma = 0$ we have that $\mathrm{gdf}(0) \approx p$. When $\gamma > \gamma^{(1)}$, i.e. we are working with a model with only the intercept term estimated by the maximum likelihood method, observing that we can write $\hat{\theta}_i(\mathbf{Y}; \gamma) = \theta(\bar{Y}) \approx \theta(\mu_i) + (\bar{Y} - \mu_i)/V(\mu_i)$ and using the first definition in equation (21) it is easy to see that $\mathrm{gdf}(\gamma) \approx 1$. When we increase the value of the parameter γ, we have a reduction of $\mathrm{gdf}(\gamma)$ since we tend to identify models with fewer variables. Finally, it is also interesting to observe that, in general, $\mathrm{gdf}(\gamma)$ is a function of the tuning parameter.

Definition (21) can be considered a natural generalization of the degrees of freedom for a linear regression model. In fact, when we consider $\hat{\boldsymbol{\mu}}^{\varphi}(\mathbf{y}) = H\mathbf{y}$, where $H = X'(X'X)^{-1}X$ is the well-known hat matrix, with canonical parameter $\hat{\theta}_i^{\varphi}(\mathbf{Y}) = \hat{\mu}_i^{\varphi}(\mathbf{Y})$ and usual dispersion $a(\phi) = \sigma^2$, then Stein's lemma (Stein, 1981) gives the exact result

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{cov}\{\hat{\mu}_i(\mathbf{Y}); Y_i\} = \sum_{i=1}^{n} \frac{\partial E_{\mathbf{Y}}\{\hat{\mu}_i(\mathbf{Y})\}}{\partial y_i} = \mathrm{tr}(H) = p, \tag{22}$$

where p are the degrees of freedom in a linear regression model.

As suggested by Efron (2004), the parametric bootstrap is the more direct way to estimate the generalized degrees of freedom (21). Park and Hastie (2007) suggested that the cardinality of the active set is a useful approximation of the generalized degrees of freedom of the L_1-penalized estimator of a GLM. Although this result is appealing, since it allows us to reduce the computational burden related to the parametric bootstrap, it is based only on the asymptotic distribution of the maximum likelihood estimators. In fact, results from the simulation study show that the cardinality of the active set is a biased estimator of the generalized degrees of freedom when we consider a logistic regression model (see Section 4.1.2). In the next subsection, we derive a useful approximation for equation (21) which can be used to compute the generalized degrees of freedom for the DGLARS method.

4.1.1. Estimating the generalized degrees of freedom

In this section, when necessary, we slightly modify our notation to emphasize the dependence of various quantities on $\mathbf{y}$.

Lemma 1. Let $\gamma \in (\gamma^{(k+1)}; \gamma^{(k)})$ be a fixed value. For any $\mathbf{y}$ there is an n-dimensional open ball with centre $\mathbf{y}$ and radius δ, denoted by $\mathcal{B}_\delta(\mathbf{y})$, such that

(a) the DGLARS estimator $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\mathbf{y}; \gamma)$ is a continuous function of $\mathbf{y}$ defined on $\mathcal{B}_\delta(\mathbf{y})$,
(b) the active set $\mathcal{A}(\mathbf{y}; \gamma)$ is locally constant with respect to $\mathbf{y}$, namely for any $\mathbf{y}^* \in \mathcal{B}_\delta(\mathbf{y})$ we have $\mathcal{A}(\mathbf{y}^*; \gamma) = \mathcal{A}(\mathbf{y}; \gamma)$, and
(c) for any $a_i \in \mathcal{A}(\mathbf{y}; \gamma)$, $\operatorname{sgn}\{r^u_{a_i}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\mathbf{y}; \gamma))\}$ is locally constant.

The proof of lemma 1 is reported in Appendix A.1. Let $\boldsymbol{\mu}^*$ be the true parameter vector; assuming that $\|\mathbf{y} - \boldsymbol{\mu}^*\| \leqslant \delta$ we have

$$\hat{\boldsymbol{\theta}}(\mathbf{y}; \gamma) = \boldsymbol{\theta}(\boldsymbol{\mu}^*; \gamma) + \frac{\partial \boldsymbol{\theta}(\boldsymbol{\mu}^*; \gamma)}{\partial \mathbf{y}}(\mathbf{y} - \boldsymbol{\mu}^*) + o(\|\mathbf{y} - \boldsymbol{\mu}^*\|), \tag{23}$$

where $\hat{\boldsymbol{\theta}}(\mathbf{y}; \gamma) = \boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\mathbf{y}; \gamma))$ and $\partial \boldsymbol{\theta}(\boldsymbol{\mu}^*; \gamma)/\partial \mathbf{y} = (\partial \theta_i(\boldsymbol{\beta}_{\mathcal{A}}(\boldsymbol{\mu}^*; \gamma))/\partial y_j)$ is the Jacobian matrix of the vector function $\boldsymbol{\theta}(\mathbf{y}; \gamma)$ evaluated at the point $\boldsymbol{\beta}_{\mathcal{A}}(\boldsymbol{\mu}^*; \gamma)$, the value of the solution curve defined by the DGLARS method and using $\boldsymbol{\mu}^*$ as response variable.

Combining equation (23) with the first element of definition (21), we obtain a first-order approximation for the generalized degrees of freedom of the DGLARS estimator:

$$\operatorname{gdf}(\gamma) \approx \frac{1}{a(\phi)} \sum_{i=1}^{n} \frac{\partial \theta_i(\boldsymbol{\mu}^*; \gamma)}{\partial y_i} \operatorname{var}(Y_i) = \sum_{i=1}^{n} \frac{\partial \mu_i(\boldsymbol{\beta}_{\mathcal{A}}(\boldsymbol{\mu}^*; \gamma))}{\partial y_i} \frac{V(\mu_i^*)}{V\{\mu_i(\boldsymbol{\beta}_{\mathcal{A}}(\boldsymbol{\mu}^*; \gamma))\}}. \tag{24}$$

To use equation (24) to define an estimator of equation (21), we need to estimate μ_i^*, for $i = 1, 2, \ldots, n$. In principle, in our setting (sparse models) we can use any penalized estimator. However, to derive all the results that we are going to see in this section, we prefer to use the maximum likelihood estimate $\mu_i(\hat{\boldsymbol{\beta}}(\mathbf{y}; 0))$. Clearly, with this choice we are supposing in the rest of this section that we are in a classical setting, namely $p < n$. Then, approximation (24) suggests the following simple estimator of the generalized degrees of freedom:

$$\widehat{\operatorname{gdf}}(\gamma) = \sum_{i=1}^{n} \frac{\partial \mu_i(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\mathbf{y}; \gamma))}{\partial y_i} \frac{V\{\mu_i(\hat{\boldsymbol{\beta}}(\mathbf{y}; 0))\}}{V\{\mu_i(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\mathbf{y}; \gamma))\}}. \tag{25}$$

Theorem 2 below gives a simplified expression of this estimator. The main problem of equation (25) is the difficulty to express the DGLARS estimates as an explicit function of $\mathbf{y}$. This problem can be overcome by using the notion of a *local co-ordinate system* that was originally proposed by Amari (1982a) and used by Wei (1998) to study the second-order asymptotic properties of

the maximum likelihood estimators in exponential dispersion models. Let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimate of $\boldsymbol{\beta}$ and consider a neighbourhood $\mathcal{O}\{\boldsymbol{\mu}(\hat{\boldsymbol{\beta}})\} \subseteq \widetilde{\mathcal{S}}$. Under the assumptions that are given in Amari (1982a), since $\mathcal{S}$ is a dually flat space, every point $\boldsymbol{\mu} \in \mathcal{O}\{\boldsymbol{\mu}(\hat{\boldsymbol{\beta}})\}$ can be specified by using the pair $\boldsymbol{\omega} = (\boldsymbol{\beta}, \boldsymbol{\zeta})$ and the notion of a -1-geodesic, namely

$$\boldsymbol{\mu}(\boldsymbol{\omega}) = \boldsymbol{\mu}(\boldsymbol{\beta}) + \mathbf{n}(\boldsymbol{\beta}, \boldsymbol{\zeta})$$

where $\boldsymbol{\zeta} \in \mathbb{R}^{n-p}$, $\mathbf{n}(\boldsymbol{\beta}, \boldsymbol{\zeta}) = \sum_{j=1}^{n-p} \mathbf{n}_j(\boldsymbol{\beta})\zeta_j$ and each $\mathbf{n}_j(\boldsymbol{\beta})$ satisfies the orthogonality condition

$$\partial_m \boldsymbol{\mu}(\boldsymbol{\beta})' I\{\boldsymbol{\mu}(\boldsymbol{\beta})\}\mathbf{n}_j(\boldsymbol{\beta}) = 0, \qquad \forall m = 1, 2, \ldots, p, \tag{26}$$

where $\partial_m \boldsymbol{\mu}(\boldsymbol{\beta}) = \partial \boldsymbol{\mu}(\boldsymbol{\beta})/\partial \beta_m$. Assuming that $\mathbf{y} \in \mathcal{O}\{\boldsymbol{\mu}(\hat{\boldsymbol{\beta}})\}$, we can specify it in the following way:

$$\mathbf{y}(\hat{\boldsymbol{\omega}}) = \mathbf{y}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}) = \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}) + \mathbf{n}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}). \tag{27}$$

Although expression (27) is obtained by using the geometrical properties of the maximum likelihood estimator, it can be modified to study the properties of the DGLARS estimator. For a fixed value $\gamma \in (\gamma^{(k+1)}; \gamma^{(k)})$, let $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma) \in \mathbb{R}^k$ be the corresponding DGLARS estimate and $\mathcal{A}(\gamma)$ the active set. Without loss of generality we shall assume that $\mathcal{B}_\delta(\mathbf{y}) \subset \mathcal{O}\{\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))\}$. Using the geometrical description that is given in Section 3 and lemma 1, it is easy to show that every $\boldsymbol{\mu} \in \mathcal{B}_\delta(\mathbf{y})$ can be specified as

$$\boldsymbol{\mu}(\boldsymbol{\omega}_{\mathcal{A}}(\gamma)) = \boldsymbol{\mu}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma)) - \gamma \mathbf{v}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma)) + \mathbf{n}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma), \boldsymbol{\zeta}),$$

where $\boldsymbol{\beta}_{\mathcal{A}}(\gamma) = \boldsymbol{\beta}_{\mathcal{A}}(\boldsymbol{\mu}; \gamma)$ is the value of the solution curve defined by using $\boldsymbol{\mu}$ as response variable, $\boldsymbol{\zeta} \in \mathbb{R}^{n-k}$ and

$$\mathbf{v}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma)) = \sum_{a_k \in \mathcal{A}(\gamma)} \partial_{a_k} \boldsymbol{\mu}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma))\, v_{a_k}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma))$$

is the vector satisfying the condition

$$\partial_{a_i} \boldsymbol{\mu}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma))' I(\boldsymbol{\mu}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma))\, \mathbf{v}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma)) = \operatorname{sgn}\{r^u_{a_i}(\gamma)\}, \qquad \forall a_i \in \mathcal{A}(\gamma). \tag{28}$$

From lemma 1, the signs of the Rao score test statistics corresponding to the active predictors are locally constant and therefore, without loss of generality, we can assume that the model is reparameterized so that $\operatorname{sgn}\{r^u_{a_i}(\gamma)\} = 1$ for any $a_i \in \mathcal{A}(\gamma)$. Like in equation (27), using the DGLARS method, the observed response vector can be specified as

$$\mathbf{y}(\hat{\boldsymbol{\omega}}_{\mathcal{A}}(\gamma)) = \mathbf{y}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma), \hat{\boldsymbol{\zeta}}) = \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)) - \gamma \mathbf{v}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)) + \mathbf{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma), \hat{\boldsymbol{\zeta}}). \tag{29}$$

In the next theorem, we propose an estimator of the generalized degrees of freedom of the DGLARS method based on specification (29).

Theorem 2. For any γ, the estimator (25) is given by

$$\begin{aligned}\widehat{\operatorname{gdf}}(\gamma) &= \operatorname{tr}\{J^{-1}_{\mathcal{A}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))\, I(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma), \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0))\} \\ &= \operatorname{tr}[\{I_{\mathcal{A}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}) + \gamma \boldsymbol{\Gamma}^1(\hat{\boldsymbol{\beta}}_{\mathcal{A}}) - \mathbf{H}^1(\hat{\boldsymbol{\omega}}_{\mathcal{A}})\}^{-1} I(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma), \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0))],\end{aligned} \tag{30}$$

where $J^{-1}_{\mathcal{A}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))$ is the observed Fisher information matrix evaluated at the point $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)$,

$$I(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma), \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)) = \frac{\partial \boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial \boldsymbol{\beta}_{\mathcal{A}}(\gamma)}' \mathbf{V}(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0))) \frac{\partial \boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial \boldsymbol{\beta}_{\mathcal{A}}(\gamma)}, \boldsymbol{\Gamma}^1(\hat{\boldsymbol{\beta}}_{\mathcal{A}}) = -\left(\frac{\partial \boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial \beta_{a_i}}' \frac{\partial \mathbf{v}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial \beta_{a_j}} \right)$$

and

$$\mathbf{H}^1(\hat{\boldsymbol{\omega}}_{\mathcal{A}}) = \left(\frac{\partial \boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial \beta_{a_i}}' \frac{\partial \mathbf{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma), \hat{\boldsymbol{\zeta}})}{\partial \beta_{a_j}} \right).$$

The proof of theorem 2 is reported in Appendix A.2. When we work with a GLM that is specified by the canonical link function, i.e. when $\mathcal{M}$ is flat with respect to the 1-connection, the elements of the matrices $\boldsymbol{\Gamma}^1(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))$ and $\mathbf{H}^1(\hat{\boldsymbol{\omega}}_{\mathcal{A}}(\gamma))$ are equal to 0, then equation (30) can be simplified.

Corollary 1. When we work with a GLM that is specified by the canonical link function, equation (30) is equal to

$$\widehat{\text{gdf}}(\gamma) = \text{tr}\{I_{\mathcal{A}}^{-1}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))\, I_{\mathcal{A}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0))\}, \tag{31}$$

where

$$I_{\mathcal{A}}(\boldsymbol{\beta}) = X'_{\mathcal{A}} \frac{\partial \boldsymbol{\mu}(\boldsymbol{\beta})}{\partial \boldsymbol{\theta}} X_{\mathcal{A}}$$

and $X_{\mathcal{A}}$ is the submatrix of design matrix X identified by the active set $\mathcal{A}(\gamma)$.

When we assume that the response vector is normally distributed and we use the canonical link function, estimator (31) is equal to the cardinality of the active set, which is the main result given in Zou *et al.* (2007). In general, the estimator proposed is different from $|\mathcal{A}(\gamma)|$. For example, when we are not working with a linear regression model and $\gamma \in (\gamma^{k+1}; \gamma^k]$ we have that the cardinality of the active set is fixed and equal to k while the proposed estimator is a real function of γ. When $\gamma = 0$, it is also interesting to note that estimator (31) is always equal to p and in this case corollary 1 gives the same result as reported in Efron (1986).

4.1.2. Generalized degrees-of-freedom estimation in logistic regression

In this section we evaluate the behaviour of the proposed estimator of the generalized degrees of freedom in a logistic regression setting. A fixed sequence of γ was used to compare estimator (31) with the cardinality of the active set. The purpose of this section is not to show the superiority of our estimator over the standard estimator, namely the cardinality of the active set. It is merely to show that both methods can exhibit bias and that in particular circumstances one should not naively rely on either. We show that in each particular case some *ad hoc* solutions may be needed to resolve these problems.

The generalized degrees of freedom were numerically evaluated; for a fixed value of γ, we independently simulated N response vectors from the GLM considered with true mean parameter vector $\boldsymbol{\mu}(\boldsymbol{\beta})$; then we compute

$$\hat{E}[\theta_i(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\mathbf{Y};\gamma))\{Y_i - \mu_i(\boldsymbol{\beta})\}] = \sum_{j=1}^{N} \frac{\theta_i^j(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\mathbf{y};\gamma))\{y_i^j - \mu_i(\boldsymbol{\beta})\}}{N}$$

Table 3. Estimated mean-squared errors of the estimators considered†

	Results for the following gdfs:										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
$\widehat{\text{gdf}}(\gamma)$	0.08	0.46	0.72	0.91	1.25	1.59	1.82	2.23	2.43	2.65	2.83
$\widehat{\lvert\mathcal{A}(\gamma)\rvert}$	0.26	1.40	2.05	2.33	2.67	2.59	2.35	2.17	1.60	1.30	1.43
$\text{gdf}_{\text{adj}}(\gamma)$	0.09	0.57	0.83	0.93	1.16	1.31	1.39	1.56	1.40	1.20	0.85

†The estimated corrector factor $\hat{\alpha}$ is approximately equal to 0.18. It was computed by using $\gamma = 0$ and 100000 bootstrap replications.

with $N = 10^6$ and we use $\Sigma_{i=1}^{n} \hat{E}[\theta_i(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)]\{Y_i - \mu_i(\boldsymbol{\beta})\}$ as an estimate of the generalized degrees of freedom ($n = 100$). This simulation study is based on a logistic regression model with 10 predictors sampled from a standard normal distribution; the predictors were centred and scaled so that for any predictor we have the Euclidean norm equal to 1. To generate the binary response vector, the coefficients are sampled from a normal distribution with expected value equal to 3 and unit variance. (We note that under this setting approximation (6.8) in Efron (1986) is not satisfied and therefore equation (21) is slightly different from p.) To compare the behaviour of the two estimators, we report the estimated mean-squared errors in Table 3 and the bias and variance in Fig. 2.

Both estimators, the raw $\widehat{\text{gdf}}(\gamma)$ and the cardinality of the active set, are biased, as shown in Fig. 2(a), whereas our proposed estimator $\widehat{\text{gdf}}(\gamma)$ seems to have a bias linearly increasing with the generalized degrees of freedom, the bias of cardinality of the active set varies in a non-linear way. The estimator $\widehat{\text{gdf}}(\gamma)$ has a slightly lower variance than the cardinality of the active

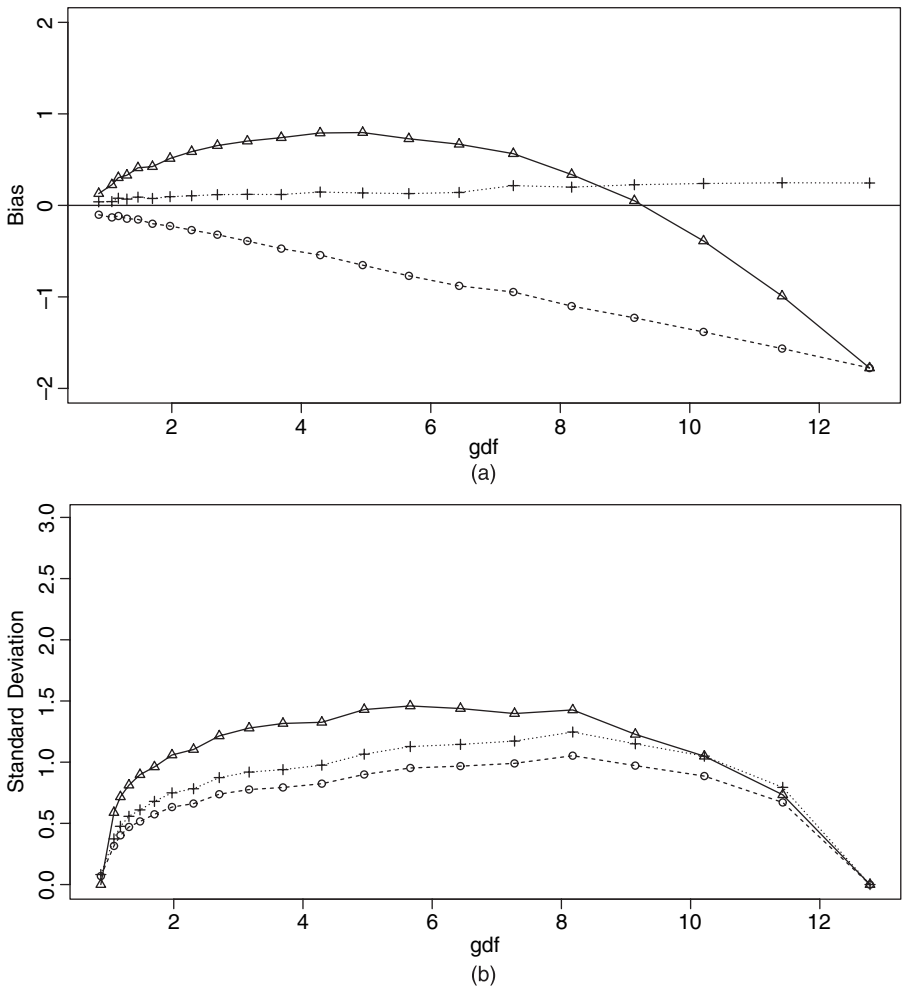

Fig. 2. Results from the simulation studies (○, $\widehat{\text{gdf}}(\gamma)$; Δ, mean card$\{\mathcal{A}(\gamma)\}$; +, $\widehat{\text{gdf}}_{\text{adj}}(\gamma)$): (a) bias and (b) standard deviation from the simulation study based on the logistic regression model

set, as shown in Fig. 2(b). Combining bias and variance, Table 3 shows that for all the cases considered there is no clear winner in terms of the mean-squared error. The simulation study suggests that a first-order approximation may not always be sufficient to obtain an unbiased estimator of $\mathrm{gdf}(\gamma)$, but it is necessary to consider a higher order expansion of equation (21). In the spirit of a Bartlett correction, the approximate linear relationship that characterizes the bias of the proposed estimator in the case of logistic regression, it is possible to consider a linearly adjusted estimator. We define the following *ad hoc* adjusted estimator of the gereralized degrees of freedom:

$$\widehat{\mathrm{gdf}}_{\mathrm{adj}}(\gamma) = \widehat{\mathrm{gdf}}(\gamma)(1+\alpha).$$

In our study, $\hat{\alpha}$ is computed by a parametric bootstrap. From the shape of the bias of the cardinality of the active set, it is clear that such a correction will not be useful for that estimator. From a computational point of view, it is important to note that this adjusted estimator is more efficient compared with the full bootstrap, since we only use the bootstrap at a single level of γ, e.g. $\gamma = 0$, merely to compute $\hat{\alpha}$. Figs 2(a) and 2(b) as well as the observed mean-squared errors suggest that the *ad hoc* adjusted estimator $\widehat{\mathrm{gdf}}_{\mathrm{adj}}$ may give slightly better results than the other estimators.

The point of this simulation study is to show that the cardinality of the active set must be used with care, but that also the generalized degrees-of-freedom estimator based on *covariance penalty theory* may be biased in some non-trivial ways. Although there may be ways to adjust for this bias, it is important to realize that model complexity in non-linear models is far from trivial. In GLM settings where the non-linearity is not too pronounced, such as Poisson regression, the cardinality of the active set is a relatively good estimate of the generalized degrees of freedom. This was observed in another simulation study based on a Poisson regression model. For brevity the results are not reported.

4.2. Relationship between differential geometric least angle regression and L_1-penalized generalized linear models

In Section 3.2 we have shown that the L_1-penalty function cannot be used to generalize the least angle regression method since it does not consider the variation that is related to the Fisher information. In this section we study the geometrical conditions under which the two methods coincide—possibly with some minor tweaking. The point is not to propose DGLARS as an algorithm for calculating the L_1-penalized GLMs—in fact, the DGLARS method may be computationally more expensive than other customized techniques. Instead, by showing in which cases the two methods vary, we can find an indication of when we might expect possible differences in performance between the two methods.

Efron *et al.* (2004) showed that the full set of lasso solutions can be obtained by a simple modification of least angle regression. Let $\hat{\boldsymbol{\beta}}$ be the solution of a GLM penalized by using the L_1-penalty function; then it is easy to show that the sign of any non-zero coefficient must agree with the sign of the score function, namely

$$\mathrm{sgn}\{\partial_m l(\hat{\boldsymbol{\beta}}; \mathbf{y})\} = \mathrm{sgn}(\hat{\beta}_m). \tag{32}$$

Like the LARS algorithm, the DGLARS method does not impose a restriction on the signs of the coefficients that are non-zero, since it is based on a generalization of the equiangularity condition. Therefore, the dimension of the active set that is identified by the DGLARS algorithm grows monotonically while the algorithm proceeds. When we use the L_1-penalty function, the dimension of the active set can shrink as the penalty parameter grows; in other

words, a covariate is removed from the active set when condition (32) is violated. This is commonly so when we are working in a high dimensional setting, as the covariates are highly correlated.

Moreover, the following theorem shows that the relationship between the DGLARS method and the L_1-penalty function is related to the link function that is used to specify the GLM.

Theorem 3. If the two conditions

(a) the solution curve defined by the L_1-penalty function satisfies the *one at a time* assumption (this means that any change in the active set when the penalty parameter changes for any L_1-penalized GLM involves only one covariate) and
(b) the Riemannian submanifold $\mathcal{M}$ is flat with respect to the Levi–Civita connection, i.e., for all l, m and n,

$$\partial_{\beta_l} i_{mn}(\boldsymbol{\beta}) = 0,$$

where $i_{mn}(\boldsymbol{\beta})$ is the generic (m, n) element of the Fisher information matrix evaluated at the point $\boldsymbol{\beta}$,

are satisfied, then the following generative *modified DGLARS algorithm* generates the solution path for an arbitrary L_1-penalized GLM whose covariates are rescaled to have zero mean and norm equal to 1. Start with $\gamma = 0$, before letting γ increase, checking the following conditions.

Step 1: if the DGLARS solution curve goes through the point $\boldsymbol{\beta}_{\mathcal{A}}(\gamma^{(k)})$ (for $k \leqslant p$), for which

$$\operatorname{sgn}\{\partial_{a_i} l(\boldsymbol{\beta}_{\mathcal{A}}(\gamma^{(k)})); \mathbf{y}\} = \operatorname{sgn}\{\beta_{a_i}(\gamma^{(k)})\}, \qquad \forall a_i \in \mathcal{A}(\gamma), \tag{33}$$

then $\boldsymbol{\beta}_{\mathcal{A}}(\gamma^{(k)})$ is also a solution obtained by using the L_1-penalty function (this condition is trivially true for the empty set $\mathcal{A} = \emptyset$ at the beginning of the DGLARS algorithm when $\gamma = 0$).
Step 2: let γ^{k*} be the first value where condition (33) is violated for covariate m^*. Then remove covariate m^* from the active set by setting $\beta_{m^*} = 0$. This is a solution of the L_1-penalized GLM.
Step 3: let γ increase, starting from $\gamma > \gamma^{k*}$, and repeat step 2 as often as needed.

See Appendix A.3 for the proof of theorem 3. The one at a time sequential nature of the DGLARS algorithm requires that at no point in the L_1-penalized GLM solution path does an infinitesimal change in the tuning parameter correspond to a change in the active set by more than one variable. This is identical to the assumption in Efron *et al.* (2004). Note that both the DGLARS and the L_1-penalized GLM can be generalized to cover more general settings, such as the case of several coefficients paths crossing zero simultaneously. We choose to avoid this generalization because it diverts from the clarity of the presentation. Condition (b) of theorem 3 is equivalent to assuming that we are working with a GLM that is specified by using the global variance stabilizing transformation as link function. In this case the Fisher information matrix satisfies the identity

$$I(\boldsymbol{\beta}) = X' \frac{\partial \boldsymbol{\mu}(\boldsymbol{\beta})}{\partial \boldsymbol{\eta}} \mathbf{V}^{-1}(\boldsymbol{\mu}(\boldsymbol{\beta})) \frac{\partial \boldsymbol{\mu}(\boldsymbol{\beta})}{\partial \boldsymbol{\eta}} X = k X' X$$

since $\partial \mu_i(\boldsymbol{\beta}) / \partial \eta_i = V^{1/2}(\mu_i(\boldsymbol{\beta}))\sqrt{k}$. For example, let us suppose that we are working with a GLM that is specified assuming that the response variable is drawn from a Poisson distribution and with variance stabilizing transformation as link function, namely $\sqrt{\mu_i(\boldsymbol{\beta})} = \mathbf{x}_i' \boldsymbol{\beta}$. In this case, it is easy to see that $\partial \mu_i(\boldsymbol{\beta}) / \partial \eta_i = 2\sqrt{\mu_i(\boldsymbol{\beta})}$. The variance stabilizing link function is usually used

in industrial applications to find orthogonal designs in GLMs. See chapter 7 in Myers *et al.* (2002) for more details.

It is possible to predict approximately where the sign violation in condition (33) may take place. Let $\mathbf{d} = D^{-1}\varphi_{\mathcal{A}}(\gamma^{(k)})\mathbf{v}_{\mathcal{A}}$; using the local approximation of the solution curve (15), it is easy to see that we would expect the next sign change for variable m to occur approximately at step size $\beta_m(\gamma^{(k)})/d_m$. The first sign change will, approximately, occur at

$$\Delta\gamma^{\text{sign}} = \min_{a_i \in \mathcal{A}} \{\beta_{a_i}(\gamma^{(k)})d_{a_i}\}.$$

If $\Delta\gamma^{\text{sign}}$ is less than the step change that is expected for the next change in the active set, $\Delta\gamma^{\text{opt}}$, then it is likely that condition (33) will be violated.

In the following section we compare DGLARS with other methods, including L_1-penalized inference, by means of simulation studies. It is important to note that DGLARS differs from L_1-regression in significant ways.

5. Simulations and practical application

5.1. Simulation study

In this section we compare DGLARS with some of the most popular sparse GLM algorithms, namely the predictor–corrector method (*glmpath*) developed by Park and Hastie (2007), the cyclical co-ordinate descent method (*glmnet*) developed by Friedman *et al.* (2010) and the gradient ascent algorithm (*penalized*) proposed in Goeman (2009, 2010).

Our simulation study is based on a logistic regression model with sample size $n = 100$ and three different values of p, namely $p \in (100, 200, 1000)$. The large values of p are useful to study the behaviour of the methods in a high dimensional setting. The study is based on five scenarios corresponding to five different configurations of the covariance structure of the p predictors. The details of the scenarios considered are as follows:

(a) $X_1, X_2, \ldots, X_p$ sampled from an $N(\mathbf{0}; \Sigma)$ distribution, where the diagonal and off-diagonal elements of Σ are 1 and 0 respectively;
(b) the same as (a) but with off-diagonal elements of Σ equal to 0.5;
(c) the same as (a) but with $\text{corr}(X_i; X_j) = \rho^{|i-j|}$ (in our study we use $\rho = 0.9$);
(d) a scenario based on a hierarchical model with two levels (let φ_1 and φ_2 be two latent variables following a standard normal distribution; the kth predictor is sampled as $X_k = \varphi_1 z_{1k} + \varphi_2 z_{2k} + z_{3k}$ where z_{ik} has a standard normal distribution);
(e) the same as (a) with Σ a block diagonal matrix, namely $\Sigma = \text{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_k)$ where Σ_i is a 10×10 matrix with diagonal elements equal to 1 and off-diagonal elements equal to 0.5.

Only the first five predictors are used to simulate the binary response variable. We choose

$$\beta = (1, \underbrace{2, \ldots, 2}_{5}, \underbrace{0, \ldots, 0}_{p-5}).$$

For DGLARS, glmnet and glmpath we simulate 1000 data sets whereas 500 data sets are simulated for the gradient ascent algorithm. A tenfold cross-validation deviance is used to select the tuning parameter for the L_1-penalized maximum likelihood estimator and the parameter γ of the method proposed. Several summary measures are used to compare the behaviour of the methods considered.

In Table 4 we report the median of the total number of variables included in the final model, the false discovery rate, the false positive rate, the false negative rate and the deviance. These results

Table 4. Results from the simulation studies; for each scenario we report the median number of variables included in the final model (Size), the false discovery rate FDR, the false positive rate FPR, the false negative rate FNR and the cross-validation deviance Dev†

Scenario	*p*	*Algorithm*	*Size*	*FDR*	*FPR*	*FNR*	*Dev*
(a)	100	DGLARS	*23.00* (0.22)	*0.77* (0.00)	*0.20* (0.00)	*1.00* (0.00)	90.68 (0.39)
		glmnet	24.00 (0.22)	0.78 (0.00)	*0.20* (0.00)	*1.00* (0.00)	82.82 (0.43)
		glmpath	26.00 (0.19)	0.80 (0.00)	0.23 (0.00)	*1.00* (0.00)	*82.65* (0.47)
		penalized	26.00 (0.29)	0.80 (0.00)	0.22 (0.00)	*1.00* (0.00)	83.18 (0.76)
	200	DGLARS	*26.00* (0.28)	*0.79* (0.00)	*0.11* (0.00)	0.98 (0.00)	97.13 (0.41)
		glmnet	26.50 (0.28)	*0.79* (0.00)	*0.11* (0.00)	0.98 (0.00)	86.67 (0.42)
		glmpath	29.00 (0.26)	0.82 (0.00)	0.12 (0.00)	0.98 (0.00)	*86.36* (0.46)
		penalized	29.00 (0.38)	0.82 (0.00)	0.12 (0.00)	*0.97* (0.00)	87.23 (0.62)
	1000	DGLARS	*29.00* (0.45)	*0.80* (0.01)	*0.02* (0.00)	0.78 (0.01)	110.15 (0.49)
		glmnet	*29.00* (0.46)	0.81 (0.01)	0.03 (0.00)	0.77 (0.01)	98.19 (0.45)
		glmpath	32.00 (0.45)	0.83 (0.00)	0.03 (0.00)	0.79 (0.01)	*97.38* (0.45)
		penalized	32.00 (0.63)	0.85 (0.00)	0.03 (0.00)	*0.73* (0.01)	98.33 (0.63)
(b)	100	DGLARS	*16.00* (0.12)	*0.71* (0.00)	*0.13* (0.00)	0.91 (0.00)	58.10 (0.33)
		glmnet	18.00 (0.12)	0.73 (0.00)	0.14 (0.00)	0.91 (0.00)	*51.69* (0.33)
		glmpath	18.00 (0.11)	0.75 (0.00)	0.15 (0.00)	0.92 (0.00)	52.36 (0.45)
		penalized	19.00 (0.15)	0.77 (0.00)	0.16 (0.00)	*0.88* (0.01)	51.83 (0.50)
	200	DGLARS	*20.00* (0.14)	*0.76* (0.00)	*0.08* (0.00)	*0.93* (0.00)	60.62 (0.31)
		glmnet	*20.00* (0.13)	0.77 (0.00)	*0.08* (0.00)	*0.93* (0.00)	51.89 (0.31)
		glmpath	21.00 (0.13)	0.78 (0.00)	0.09 (0.00)	0.94 (0.00)	*51.73* (0.33)
		penalized	29.00 (0.38)	0.82 (0.00)	0.12 (0.00)	0.97 (0.00)	87.23 (0.62)
	1000	DGLARS	*25.00* (0.16)	0.87 (0.00)	*0.02* (0.00)	*0.62* (0.01)	68.53 (0.32)
		glmnet	26.00 (0.13)	0.86 (0.00)	*0.02* (0.00)	0.71 (0.01)	*56.69* (0.30)
		glmpath	27.00 (0.13)	0.86 (0.00)	*0.02* (0.00)	0.72 (0.01)	57.09 (0.36)
		penalized	32.00 (0.63)	*0.85* (0.00)	0.03 (0.00)	0.73 (0.01)	98.33 (0.63)
(c)	100	DGLARS	*9.00* (0.13)	*0.47* (0.01)	*0.05* (0.00)	0.83 (0.00)	45.29 (0.32)
		glmnet	12.00 (0.13)	0.63 (0.00)	0.09 (0.00)	*0.81* (0.00)	*42.99* (0.31)
		glmpath	14.00 (0.12)	0.69 (0.00)	0.10 (0.00)	0.82 (0.00)	44.11 (0.46)
		penalized	13.00 (0.19)	0.67 (0.01)	0.10 (0.00)	*0.81* (0.01)	44.83 (0.66)
	200	DGLARS	*10.00* (0.15)	*0.54* (0.01)	*0.03* (0.00)	0.79 (0.01)	48.55 (0.33)
		glmnet	15.00 (0.17)	0.70 (0.00)	0.06 (0.00)	*0.78* (0.00)	*46.68* (0.34)
		glmpath	16.00 (0.15)	0.75 (0.00)	0.07 (0.00)	*0.78* (0.00)	46.81 (0.40)
		penalized	16.00 (0.22)	0.74 (0.00)	0.07 (0.00)	0.79 (0.01)	47.85 (0.64)
	1000	DGLARS	*12.00* (0.19)	*0.63* (0.01)	*0.01* (0.00)	*0.73* (0.01)	*43.46* (0.30)
		glmnet	21.00 (0.25)	0.77 (0.00)	0.02 (0.00)	0.78 (0.00)	44.05 (0.32)
		glmpath	22.00 (0.24)	0.80 (0.00)	0.02 (0.00)	0.78 (0.00)	45.25 (0.43)
		penalized	23.00 (0.33)	0.80 (0.00)	0.02 (0.00)	0.80 (0.01)	45.72 (0.63)
(d)	100	DGLARS	*22.00* (0.19)	*0.75* (0.00)	*0.18* (0.00)	*1.00* (0.00)	64.87 (0.33)
		glmnet	23.00 (0.19)	0.76 (0.00)	0.19 (0.00)	*1.00* (0.00)	55.41 (0.30)
		glmpath	24.00 (0.17)	0.79 (0.00)	0.21 (0.00)	*1.00* (0.00)	54.97 (0.33)
		penalized	24.00 (0.24)	0.78 (0.00)	0.20 (0.00)	*1.00* (0.00)	*54.70* (0.46)
	200	DGLARS	*29.00* (0.25)	*0.81* (0.00)	*0.12* (0.00)	*1.00* (0.00)	68.51 (0.41)
		glmnet	29.50 (0.25)	0.82 (0.00)	0.13 (0.00)	*1.00* (0.00)	59.34 (0.35)
		glmpath	31.00 (0.23)	0.83 (0.00)	0.13 (0.00)	*1.00* (0.00)	59.31 (0.40)
		penalized	31.00 (0.32)	0.83 (0.00)	0.13 (0.00)	*1.00* (0.00)	*58.70* (0.59)
	1000	DGLARS	*26.00* (0.34)	*0.79* (0.00)	*0.02* (0.00)	0.90 (0.00)	106.51 (0.46)
		glmnet	*26.00* (0.36)	0.80 (0.00)	*0.02* (0.00)	*0.85* (0.00)	88.11 (0.43)
		glmpath	29.00 (0.34)	0.82 (0.00)	*0.02* (0.00)	0.87 (0.00)	86.36 (0.41)
		penalized	32.00 (0.46)	0.84 (0.00)	0.03 (0.00)	0.90 (0.01)	*84.14* (0.50)
(e)	100	DGLARS	*15.00* (0.16)	*0.65* (0.00)	*0.11* (0.00)	*0.98* (0.00)	63.86 (0.31)
		glmnet	18.00 (0.18)	0.71 (0.00)	0.14 (0.00)	*0.98* (0.00)	60.55 (0.35)
		glmpath	20.00 (0.16)	0.74 (0.00)	0.16 (0.00)	*0.98* (0.00)	60.29 (0.40)
		penalized	19.00 (0.24)	0.73 (0.00)	0.15 (0.00)	*0.98* (0.00)	*54.79* (0.52)

(*continued*)

Table 4 (*continued*)

Scenario	*p*	*Algorithm*	*Size*	*FDR*	*FPR*	*FNR*	*Dev*
(e)	200	DGLARS	*15.00* (0.18)	*0.64* (0.00)	*0.05* (0.00)	*0.97* (0.00)	57.38 (0.30)
		glmnet	18.00 (0.20)	0.70 (0.00)	0.07 (0.00)	*0.97* (0.00)	53.49 (0.31)
		glmpath	20.00 (0.18)	0.74 (0.00)	0.08 (0.00)	*0.97* (0.00)	53.18 (0.36)
		penalized	23.00 (0.26)	0.77 (0.00)	0.09 (0.00)	0.99 (0.00)	*52.04* (0.56)
	1000	DGLARS	*22.00* (0.29)	*0.73* (0.00)	*0.02* (0.00)	0.97 (0.00)	64.43 (0.41)
		glmnet	27.00 (0.34)	0.77 (0.00)	*0.02* (0.00)	*0.95* (0.00)	61.68 (0.40)
		glmpath	29.00 (0.31)	0.80 (0.00)	*0.02* (0.00)	*0.95* (0.00)	62.05 (0.52)
		penalized	29.00 (0.37)	0.82 (0.00)	*0.02* (0.00)	0.96 (0.00)	*52.23* (0.56)

†Standard errors are in parentheses. Italic values identify the best methods for each scenario.

show that the global behaviour of the DGLARS method is closely related to the covariance structure among the p predictors. When we consider scenario (a), DGLARS exhibits similar behaviour to the co-ordinate descent method that was proposed by Friedman *et al.* (2010), whereas the algorithm that was developed by Goeman (2009, 2010) gives similar results to the path following algorithm that was developed by Park and Hastie (2007). Table 4 also shows that glmpath achieves a higher proportion of models including the true variables; however, it achieves this by also including a larger number of falsely selected variables. The gradient ascent algorithm and glmpath usually tend to select larger models. When the predictors are highly correlated as in scenario (b), the algorithm that was proposed by Goeman (2009, 2010) tends to be characterized by a low proportion of models including all the true predictors. Like in scenario (a), glmpath tends to identify larger models and with a larger number of falsely selected variables. When the relevant predictors are highly correlated (scenario (c)), or when there is an unknown group structure of the predictors (scenario (e)), the results show that the DGLARS algorithm selects sparser models than the other methods. This feature is also associated with a lower proportion of false variables included in the final models and with a higher specificity. Similar results are obtained in scenario (d).

5.2. Application of differential geometric least angle regression to a genomics data set

In this section we demonstrate the use of the proposed method in a logistic regression model applied to a breast cancer gene deletion–amplification data set that was obtained by John Bartlett at the Royal Infirmary, Glasgow (Wit and McClure, 2004). The aim of the study is to identify which genes play a crucial role in the severity of the disease, defined as whether or not the patient dies as a result of breast cancer. Excluding 10 samples for which there is no survival information, the data set contains 105 patients, 47 of whom are labelled as deceased through breast cancer.

Table 5. Results from the breast cancer data

Method	*Testing error*	*Number of genes*
DGLARS algorithm	15/53	2
L_1-penalized logistic regression	17/53	9

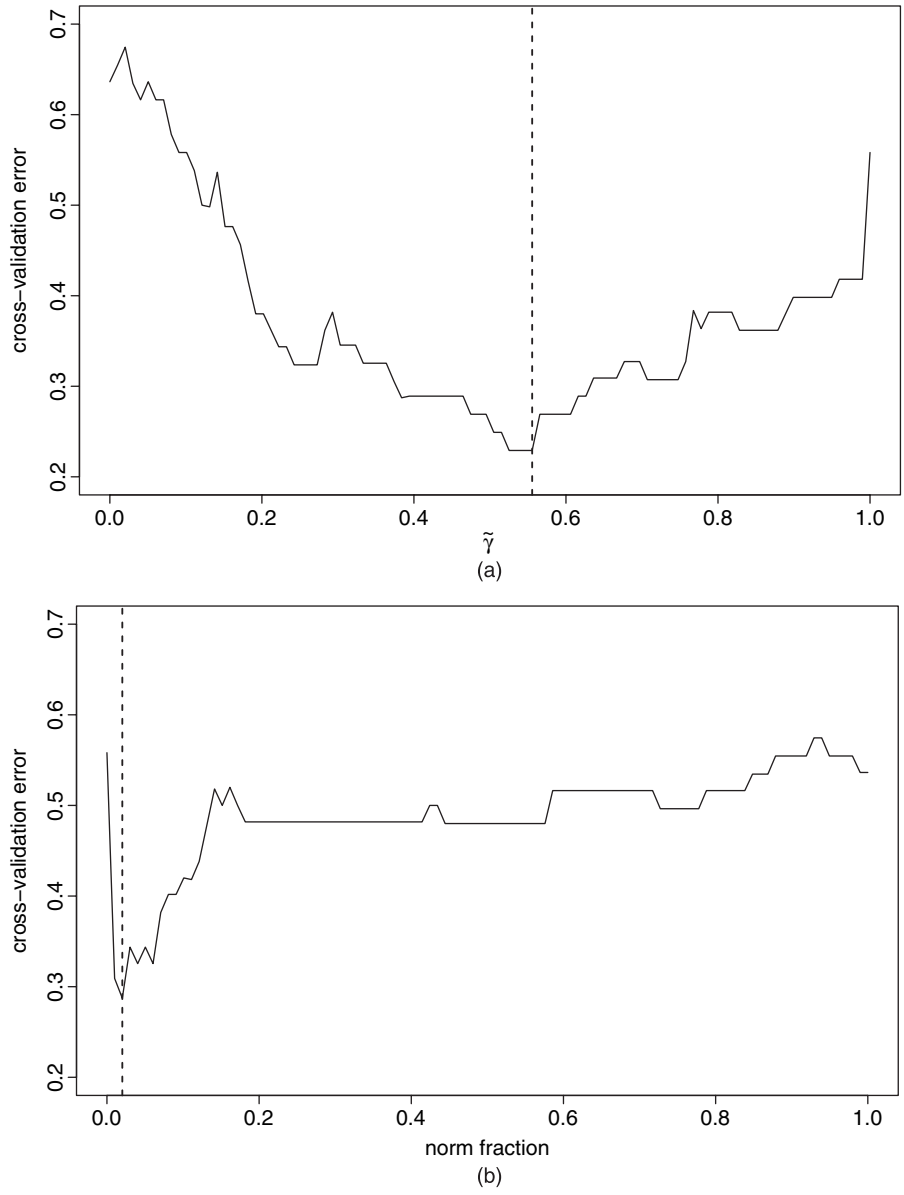

Fig. 3. (a) Cross-validation errors computed for the DGLARS method and (b) for the L_1-penalized logistic regression model: for DGLARS, the path is computed as a function of the parameter $\tilde{\gamma}$, the ratio of γ and the maximum value of γ; for the L_1-penalized logistic regression model, the path is computed as a function of the fraction of the L_1-norm (⋮, selected levels of regularization for the two methods)

For each patient, 287 gene deletion–amplification measurements are available. A few missing covariate values are imputed by using the method that was proposed by Troyanskaya *et al.* (2001).

We randomly select 52 patients, of whom 29 patients are labelled as having died through breast cancer, to be used as the training set. The remaining 53 patients are used as the test set. The tuning parameter of the L_1-penalized logistic regression model and the parameter γ of

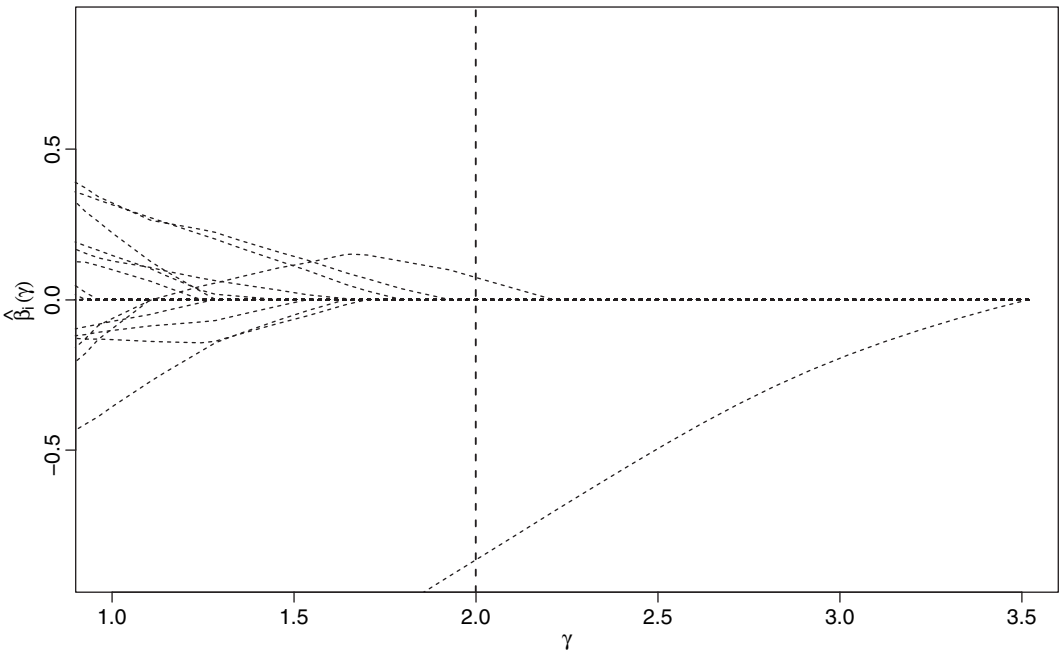

Fig. 4. Path of the coefficients as a function of the parameter γ: $^{|}$, value of γ selected by fivefold cross-validation

the DGLARS method are obtained by using fivefold cross-validation. Fig. 3 shows the dependence of the cross-validation errors computed on the training set on the tuning parameters for both methods. In Table 5 we report the test error and the number of genes used to define the classification rules. From Table 5 we can see that the DGLARS method is marginally more accurate than the L_1-penalized logistic regression model; the test error is equal to 15/53 for the DGLARS and 17/53 for the L_1-penalized logistic regression model. This analysis confirms one of the main results that was obtained in the simulation study: DGLARS is characterized by the ability to identify a sparser model than the L_1-penalty function. Only two genes are used for the classification rule that is defined by the DGLARS method, whereas the L_1-penalty function defines a classification rule based on nine genes.

The genes that were used in the DGLARS classification rule are PTGS2(COX-2) and SHGC4-207. Interestingly, it is known that the expression of COX-2 is upregulated in many cancers. Furthermore, the product of COX-2, PGH2, is converted by prostaglandin E2 synthase into PGE2 which in turn can stimulate cancer progression. Consequently it has been reported in Menter *et al.* (2010) that inhibiting COX-2 may benefit the prevention and treatment of these types of cancer. About SHGC4-207 much less is known and it could be an interesting candidate for further follow-up. Fig. 4 shows the solution path that was identified by the DGLARS method. The vertical line identifies the value of γ that was selected by fivefold cross-validation.

6. Discussion

In this paper we have proposed a new method to select important variables in a GLM. Our method is based on the geometrical structure underlying the GLM which allows us to use the signed Rao score test statistic to define a genuine generalization of the equiangularity condition

that was proposed by Efron *et al.* (2004) for linear models. Whereas LARS originally was introduced as a way to simplify L_1-penalized regression, in this paper we showed that our method gives us a more direct way to connect the geometry of the model to the sparsity of the feature space. Important theoretical questions, such as consistency of the method, are still open and will be addressed in a future work. Nevertheless, simulation studies involving the method and its relationship to sure independent screening (Fan and Song, 2010) make us hopeful about consistency. Another interesting question is related to possible connections of the method to the generalized Dantzig selector (James and Radchenko, 2009), defined as solution of the minimization problem

$$\min\|\beta\|_1 \quad \text{subject to} \quad |\partial_m l(\beta;\mathbf{y})| \leqslant \lambda \qquad (m=1,\ldots,p),$$

where $\|\cdot\|_1$ is the L_1-norm. For the linear regression model, James *et al.* (2009) provided general conditions on the design matrix under which a given lasso solution is identical to the corresponding Dantzig selector solution.

In this paper we also addressed the problem of how to define the degrees of freedom of the DGLARS method proposed. Although covariance penalty theory provides us with a general framework, we have only been able to derive a first-order approximation explicitly, by combining using the dual structure of the exponential family. A simulation study seems to show that, in some cases, the behaviour of the estimator proposed can be further improved by using a Bartlett-like correction factor. Again, the theoretical foundation of the correction factor and how to estimate it will have to be addressed in future work.

Although the current version of the DGLARS method is based on the geometrical structure of a GLM, our approach can be extended to other models, such as Cox proportional hazard models and regression models based on the quasi-likelihood function. Another important feature of the method proposed is that it can be easily extended to deal with factorial models, something that is not trivial with L_1-penalized regression. This suggests that the approach that is proposed in this paper has further mileage.

Appendix A: Proofs

A.1. Proof of lemma 1

Let $\gamma \in (\gamma^{(k+1)};\gamma^{(k)})$ be a fixed value and let $\hat{\beta}_{\mathcal{A}}(\mathbf{y};\gamma)$ be the estimate given by the DGLARS method, namely, $\hat{\beta}_{\mathcal{A}}(\mathbf{y};\gamma)$ solves the system of $k+1$ non-linear equations

$$\begin{aligned}
\partial_{a_0} l(\hat{\beta}_{\mathcal{A}}(\mathbf{y};\gamma);\mathbf{y}) &= 0,\\
r^u_{a_1}(\hat{\beta}_{\mathcal{A}}(\mathbf{y};\gamma);\mathbf{y}) - v_{a_1}\gamma &= 0,\\
\vdots \qquad & \qquad \vdots\\
r^u_{a_k}(\hat{\beta}_{\mathcal{A}}(\mathbf{y};\gamma);\mathbf{y}) - v_{a_k}\gamma &= 0.
\end{aligned}$$

As a consequence of the inverse function theorem, there is an n-dimensional open ball with centre $\mathbf{y}$ and radius $\delta_{\mathcal{A}}$, denoted by $\mathcal{B}_{\delta_{\mathcal{A}}}(\mathbf{y})$, such that $\hat{\beta}_{\mathcal{A}}(\mathbf{y};\gamma)$ is a continuous function, with respect to $\mathbf{y}$, defined on $\mathcal{B}_{\delta_{\mathcal{A}}}(\mathbf{y})$. Let $r^u_{a^c_j}(\hat{\beta}_{\mathcal{A}}(\mathbf{y};\gamma))$ be the Rao score test corresponding to the jth predictor belonging to the complement of the active set $\mathcal{A}(\mathbf{y};\gamma)$ denoted by $A^c(\mathbf{y};\gamma)$. For any $a^c_j \in \mathcal{A}^c(\mathbf{y};\gamma)$, since $r^u_{a^c_j}(\hat{\beta}_{\mathcal{A}}(\mathbf{y};\gamma);\mathbf{y})$ is a continuous function on $\mathcal{B}_{\delta_{\mathcal{A}}}(\mathbf{y})$, we have that there is a $\delta_{a^c_j} > 0$ such that $|r^u_{a^c_j}(\hat{\beta}_{\mathcal{A}}(\mathbf{y}^*;\gamma);\mathbf{y}^*)| - \gamma$ is negative for any $\mathbf{y}^* \in \mathcal{B}_{\delta_{a^c_j}}(\mathbf{y})$.

Let $\delta = \min\{\delta_{\mathcal{A}}, \delta_{a^c_1}, \ldots, \delta_{a^c_{p-k}}\}$ and consider the open ball $\mathcal{B}_{\delta}(\mathbf{y})$; then for any $\mathbf{y}^* \in \mathcal{B}_{\delta}(\mathbf{y})$ we have

$$\begin{aligned}
r^u_{a_i}(\hat{\beta}_{\mathcal{A}}(\mathbf{y}^*;\gamma);\mathbf{y}^*) &= v_{a_i}\gamma, && \forall a_i \in \mathcal{A}(\mathbf{y};\gamma),\\
|r^u_{a^c_j}(\hat{\beta}_{\mathcal{A}}(\mathbf{y}^*;\gamma);\mathbf{y}^*)| &< \gamma, && \forall a^c_j \in \mathcal{A}^c(\mathbf{y};\gamma),
\end{aligned}$$

which means that $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\mathbf{y}^*;\gamma)$ is the DGLARS estimator and $\mathcal{A}(\mathbf{y}^*;\gamma)=\mathcal{A}(\mathbf{y};\gamma)$, namely the active set is locally constant. The local constancy of $\operatorname{sgn}\{r^u_{a_i}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\mathbf{y};\gamma))\}$ follows by the continuity of $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\mathbf{y};\gamma)$.

A.2. Proof of theorem 2

The proof of theorem 2 is based on a generalization of lemma 3.2 in Kato (2009) to a more general setting. Using our lemma 1 and expression (29), we have

$$\frac{\partial\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\mathbf{y};\gamma))}{\partial\mathbf{y}}=\frac{\partial\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\omega}_{\mathcal{A}}(\gamma)}\left(\frac{\partial\mathbf{y}(\hat{\boldsymbol{\omega}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\omega}_{\mathcal{A}}(\gamma)}\right)^{-1}=\left(\frac{\partial\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\beta}_{\mathcal{A}}(\gamma)}\quad\mathbf{0}\right)\left(\frac{\partial\mathbf{y}(\hat{\boldsymbol{\omega}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\omega}_{\mathcal{A}}(\gamma)}\right)^{-1},\tag{34}$$

where $\mathbf{0}$ is an $n\times(n-|\mathcal{A}(\gamma)|)$ matrix with elements equal to 0 and

$$\begin{aligned}\frac{\partial\mathbf{y}(\hat{\boldsymbol{\omega}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\omega}_{\mathcal{A}}(\gamma)}&=\left(\frac{\partial\mathbf{y}(\hat{\boldsymbol{\omega}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\beta}_{\mathcal{A}}(\gamma)}\quad\frac{\partial\mathbf{y}(\hat{\boldsymbol{\omega}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\zeta}}\right)\\&=\left(\frac{\partial\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\beta}_{\mathcal{A}}(\gamma)}-\gamma\frac{\partial\mathbf{v}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\beta}_{\mathcal{A}}(\gamma)}+\frac{\partial\mathbf{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma);\hat{\boldsymbol{\zeta}})}{\partial\boldsymbol{\beta}_{\mathcal{A}}(\gamma)}\quad\mathbf{N}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))\right),\end{aligned}$$

where $\mathbf{N}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))$ is the orthonormal matrix with jth column equal to $\mathbf{n}_j(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))$. Using the QR-decomposition with respect to the Fisher information matrix we can write

$$\frac{\partial\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\beta}_{\mathcal{A}}(\gamma)}=\mathbf{B}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))\mathbf{R}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)),$$

where $\mathbf{B}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))$ is an $n\times|\mathcal{A}(\gamma)|$ orthonormal matrix and $\mathbf{R}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))$ is an $|\mathcal{A}(\gamma)|\times|\mathcal{A}(\gamma)|$ upper triangular matrix. Then we have

$$(\mathbf{B}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))\quad\mathbf{N}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)))'\frac{\partial\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\mu}}\frac{\partial\mathbf{y}(\hat{\boldsymbol{\omega}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\omega}_{\mathcal{A}}(\gamma)}=\begin{pmatrix}\mathbf{A}_{11}&\mathbf{0}\\\mathbf{A}_{21}&\mathbf{I}\end{pmatrix},\tag{35}$$

where

$$\mathbf{A}_{11}=\mathbf{B}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))'\frac{\partial\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\mu}}\frac{\partial\mathbf{y}(\hat{\boldsymbol{\omega}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\beta}_{\mathcal{A}}(\gamma)}$$

and

$$\mathbf{A}_{12}=\mathbf{N}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))'\frac{\partial\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\mu}}\frac{\partial\mathbf{y}(\hat{\boldsymbol{\omega}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\beta}_{\mathcal{A}}(\gamma)}.$$

Using equation (35) and the standard formula for the inverse of a partitioned matrix, we have

$$\left(\frac{\partial\mathbf{y}(\hat{\boldsymbol{\omega}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\omega}_{\mathcal{A}}(\gamma)}\right)^{-1}=\begin{pmatrix}\mathbf{A}_{11}^{-1}&\mathbf{0}\\-\mathbf{A}_{21}\mathbf{A}_{11}^{-1}&\mathbf{I}\end{pmatrix}(\mathbf{B}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))\quad\mathbf{N}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)))'\frac{\partial\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\mu}};$$

then, using equation (34) and simple algebra it is easy to show that

$$\frac{\partial\hat{\boldsymbol{\mu}}(\mathbf{y};\gamma)}{\partial\mathbf{y}}=\frac{\partial\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\beta}_{\mathcal{A}}(\gamma)}\left(\frac{\partial\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))'}{\partial\boldsymbol{\beta}_{\mathcal{A}}(\gamma)}\frac{\partial\mathbf{y}(\hat{\boldsymbol{\omega}}_{\mathcal{A}}(\gamma))}{\partial\boldsymbol{\beta}_{\mathcal{A}}(\gamma)}\right)^{-1}\frac{\partial\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))'}{\partial\boldsymbol{\beta}_{\mathcal{A}}(\gamma)}.$$

Observing that condition (28) can be written as

$$\partial_{a_i}\boldsymbol{\theta}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma))'\mathbf{v}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma))=1,\qquad\forall a_i\in\mathcal{A}(\gamma),$$

where $\partial_{a_i}\boldsymbol{\theta}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma))=\partial\boldsymbol{\theta}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma))/\partial\beta_{a_i}(\gamma)$, then taking the derivative with respect to $\beta_{a_j}(\gamma)$ of the previous identity we obtain

$$-\partial_{a_i}\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))'\,\partial_{a_j}\mathbf{v}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))=\partial_{a_ia_j}\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))'\mathbf{v}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)),\tag{36}$$

where $\partial_{a_j}\mathbf{v}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))=\partial\mathbf{v}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))/\partial\beta_{a_j}(\gamma)$ and the same notation is used for $\partial_{a_ia_j}\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))$. Similarly, condition (26) can be written as

$$\partial_{a_i}\boldsymbol{\theta}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma))'\mathbf{n}(\boldsymbol{\beta}_{\mathcal{A}}(\gamma))=0,\qquad\forall a_i\in\mathcal{A}(\gamma);$$

then taking the derivative with respect to $\beta_{a_j}(\gamma)$ we have the identity

$$-\partial_{a_i}\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))'\,\partial_{a_j}\mathbf{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))=\partial_{a_ia_j}\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))'\mathbf{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)).\tag{37}$$

Using identities (36) and (37) we have that the element (a_i, a_j) of the matrix

$$\frac{\partial \boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial \boldsymbol{\beta}_{\mathcal{A}}(\gamma)}' \frac{\partial \mathbf{y}(\hat{\boldsymbol{\omega}}_{\mathcal{A}}(\gamma))}{\partial \boldsymbol{\beta}_{\mathcal{A}}(\gamma)}$$

is equal to

$$i_{a_i a_j}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)) - \partial_{a_i a_j}\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))'\{\mathbf{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}(\gamma))\} = -\partial_{a_i a_j} l(\hat{\boldsymbol{\beta}}(\gamma); \mathbf{y}),$$

where $i_{a_i a_j}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))$ is the element (a_i, a_j) of the Fisher information matrix evaluated at the point $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)$. The previous identity shows that

$$\frac{\partial \boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial \boldsymbol{\beta}_{\mathcal{A}}(\gamma)}' \frac{\partial \mathbf{y}(\hat{\boldsymbol{\omega}}_{\mathcal{A}}(\gamma))}{\partial \boldsymbol{\beta}_{\mathcal{A}}(\gamma)}$$

is the observed Fisher information matrix evaluated at the point $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)$, which we denote as $J_{\mathcal{A}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))$. Assumptions in Section 2 tell us that $J_{\mathcal{A}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))$ is a positive definite matrix; then, using the previous results, estimator (25) is equal to

$$\begin{aligned}
\widehat{\mathrm{gdf}}(\gamma) &= \mathrm{tr}\left\{ \mathbf{V}(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0))) \frac{\partial \boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial \boldsymbol{\mu}} \frac{\partial \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial \boldsymbol{\beta}_{\mathcal{A}}(\gamma)} J_{\mathcal{A}}^{-1}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)) \frac{\partial \boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial \boldsymbol{\beta}_{\mathcal{A}}(\gamma)}' \right\} \\
&= \mathrm{tr}\left\{ \mathbf{V}(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0))) \frac{\partial \boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial \boldsymbol{\beta}_{\mathcal{A}}(\gamma)} J_{\mathcal{A}}^{-1}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)) \frac{\partial \boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial \boldsymbol{\beta}_{\mathcal{A}}(\gamma)}' \right\} \\
&= \mathrm{tr}\left\{ J_{\mathcal{A}}^{-1}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)) \frac{\partial \boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial \boldsymbol{\beta}_{\mathcal{A}}(\gamma)}' \mathbf{V}(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0))) \frac{\partial \boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\partial \boldsymbol{\beta}_{\mathcal{A}}(\gamma)} \right\} \\
&= \mathrm{tr}\{ J_{\mathcal{A}}^{-1}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))\, I(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma), \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)) \}.
\end{aligned}$$

The second equation of expression (30) is obtained observing that

$$\partial_{a_i a_j}\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))'\mathbf{v}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)) = \sum_{a_k \in \mathcal{A}} \Gamma^1_{a_i a_j a_k}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))\, v_{a_k}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)),$$

$$\partial_{a_i a_j}\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))'\mathbf{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)) = \sum_{a_k^{\mathrm{c}} \in \mathcal{A}^{\mathrm{c}}(\gamma)} H^1_{a_i a_j a_k^{\mathrm{c}}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))\hat{\zeta}_{a_k^{\mathrm{c}}}$$

where $\Gamma^1_{a_j a_j a_k}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))$ is the exponential connection of the curved exponential family (see equation (2.10) in Amari (1982a)) and $H^1_{a_i a_j a_k^{\mathrm{c}}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))$ is the tensor which defines the 1-curvature of a curved exponential family (see equation (2.9) in Amari (1982a)).

A.3. Proof of theorem 3

The Riemannian submanifold $\mathcal{M}$ is flat with respect to the Levi–Civita connection when, for any $\boldsymbol{\beta} \in \mathbb{R}^p$, the following condition is satisfied:

$$\partial_{\beta_l} i_{mn}(\boldsymbol{\beta}) = 0, \qquad \forall l, m, n. \tag{38}$$

Condition (38) implies that the link function is the variance stabilizing transformation. In that case, for any $a_i \in \mathcal{A}$, we have $i_{a_i}(\boldsymbol{\beta}) = c^2 \|\mathbf{x}_{a_i}\|^2$ (see Atkinson and Mitchell (1981) for more general results). When we work with a linear regression model it is easy to show that $c = 1$, whereas for a Poisson regression model we have $c = 2$. Without loss of generality we assume that $\|\mathbf{x}_{a_i}\| = c^{-1}$. When condition (38) is satisfied, system (13) is equal to

$$\left.\begin{aligned}
\partial_{a_0} l(\boldsymbol{\beta}(\gamma)) &= 0, \\
\partial_{a_1} l(\boldsymbol{\beta}(\gamma)) &= \mathrm{sgn}\{\partial_{a_1} l(\boldsymbol{\beta}(\gamma))\}\gamma, \\
&\vdots \qquad\qquad \vdots \\
\partial_{a_k} l(\boldsymbol{\beta}(\gamma)) &= \mathrm{sgn}\{\partial_{a_k} l(\boldsymbol{\beta}(\gamma))\}\gamma.
\end{aligned}\right\} \tag{39}$$

System (39) tells us that the algorithm proposed identifies the solution curve that is defined using a GLM with L_1-penalty function, if there is an interval $\Gamma \subseteq (\gamma^{(k+1)}; \gamma^{(k)}]$ such that

$$\text{sgn}\{\partial_{a_i} l(\boldsymbol{\beta}(\gamma))\} = \text{sgn}\{\beta_{a_i}(\gamma)\}, \qquad \forall \gamma \in \Gamma \text{ and } \forall a_i \in \mathcal{A}. \tag{40}$$

Condition (40) shows that the modified DGLARS and the L_1-penalized GLM have the same solution curve.

References

Allgower, E. and Georg, K. (2003) *Introduction to Numerical Continuation Methods*. New York: Society for Industrial and Applied Mathematics.

Amari, S.-I. (1982a) Geometrical theory of asymptotic ancillarity and conditional inference. *Biometrika*, **67**, 1–17.

Amari, S.-I. (1982b) Differential geometry of curved exponential families-curvatures and information loss. *Ann. Statist.*, **10**, 357–385.

Amari, S.-I. (1985) *Differential-geometrical Methods in Statistics*. New York: Springer.

Amari, S.-I. and Nagaoka, H. (2000) *Methods of Information Geometry*. Providence: American Mathematical Society.

Atkinson, C. and Mitchell, A. F. S. (1981) Rao's distance measure. *Sankhya* A, **43**, 345–365.

Bregman, L. M. (1967) The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Phys.*, **7**, 200–217.

Burbea J. and Rao, R. C. (1982) Entropy differential metric, distance and divergence measures in probability spaces—a unified approach. *J. Multiv. Anal.*, **12**, 575–596.

Burnham, K. P. and Anderson, D. R. (2000) *Model Selection and Inference: a Practical Information-theoretical Approach*. New York: Springer.

Candes, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger that n (with discussion). *Ann. Statist.*, **35**, 2313–2351.

do Carmo, M. P. (1992) *Riemannian Geometry*. Boston: Birkhäuser.

Efron, B. (1986) How biased is the apparent error rate of a prediction rule? *J. Am. Statist. Ass.*, **81**, 461–470.

Efron, B. (2004) The estimation of prediction error: covariance penalties and cross-validation. *J. Am. Statist. Ass.*, **99**, 619–632.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression (with discussion). *Ann. Statist.*, **32**, 407–451.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Fan, J. and Song, R. (2010) Sure independence screening in generalized linear models with *NP*-dimensionality. *Ann. Statist.*, **38**, 3567–3604.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softwr.*, **33**, 1–22.

Goeman, J. (2009) **penalized**: L1 (lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model. *R Package Version 0.9-32*. University Medical Center, Leiden. (Available from `http://cran.r-project.org/web/packages/penalized/index.html`.)

Goeman, J. (2010) L1 penalized estimation in the Cox proportional hazards model. *Biometr. J.*, **52**, 70–84.

Hesterberg, T., Choi, N. H., Meire, L. and Fraley, C. (2008) Least angle and l_1 penalized regression: a review. *Statist. Surv.*, **2**, 61–93.

Hocking, R. R. (1976) The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1–49.

Huang, J. and Zhang, T. (2010) The benefit of group sparsity. *Ann. Statist.*, **38**, 1978–2004.

James, G. M. (2002) Generalized linear models with functional predictors. *J. R. Statist. Soc.* B, **64**, 411–432.

James, G. M. and Radchenko, P. (2009) A generalized Dantzig selector with shrinkage tuning. *Biometrika*, **96**, 323–337.

James, G. M., Radchenko, P. and Lv, J. (2009) DASSO: connections between the Dantzig selector and lasso. *J. R. Statist. Soc.* B, **71**, 127–142.

Jolliffe, I. T. (1982) A note on the use of principal components in regression. *Appl. Statist.*, **31**, 300–303.

Kass, R. and Vos, P. W. (1997) *Geometrical Foundation of Asymptotic Inference*. New York: Wiley.

Kato, K. (2009) On the degrees of freedom in shrinkage estimation. *J. Multiv. Anal.*, **100**, 1338–1352.

Li, Y., Wang, N. and Carroll, R. J. (2010) Generalized functional linear models with semi-parametric single-index interaction. *J. Am. Statist. Ass.*, **105**, 621–633.

Madigan, D. and Ridgeway, G. (2004) Discussion to least angle regression. *Ann. Statist.*, **32**, 465–469.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Meier, L., van de Geer, S. and Bühlmann, P. (2009) High-dimensional additive modelling. *Ann. Statist.*, **37**, 3779–3821.

Menter, D. G., Schilsky, R. L. and Dubois, R. N. (2010) Cyclooxygenase-2 and cancer treatment: understanding the risk should be worth the reward. *Clin. Cancer Res.*, **16**, 1384–1390.

Müller, H.-G. and Stadtmüller, U. (2005) Generalized functional linear models. *Ann. Statist.*, **33**, 774–805.
Myers, R. H., Montgomery, D. C. and Vining, G. G. (2007) *Generalized Linear Models: with Applications in Engineering and the Sciences*. New York: Wiley.
Park, M. Y. and Hastie, T. (2007) L_1-regularization path algorithm for generalized linear models. *J. R. Statist. Soc.* B, **69**, 659–677.
Rao, C. R. (1945) Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calc. Math. Soc.*, **37**, 81–91.
Shen, X., Huang, H.-C. and Ye, J. (2004) Adaptive model selection and assessment for exponential family distributions. *Technometrics*, **46**, 306–317.
Spivak, M. (1979) *A Comprehensive Introduction to Differential Geometry*, 2nd edn. Boston: Publish or Perish.
Stein, C. (1981) Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, **9**, 1135–1151.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001) Missing value estimation methods for DNA. *Bioinformatics*, **17**, 520–525.
Vos, P. W. (1991) A geometric approach to detecting influential cases. *Ann. Statist.*, **19**, 1570–1581.
Wei, B.-C. (1998) *Exponential Family Nonlinear Models*. Singapore: Springer.
Wit, E. C. and McClure, J. D. (2004) *Statistics for Microarrays: Design, Analysis and Inference*. Chichester: Wiley.
Wu, T. T. and Lange, K. (2008) Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Statist.*, **2**, 224–244.
Ye, J. (1998) On measuring and correcting the effects of data mining and model selection. *J. Am. Statist. Ass.*, **93**, 120–131.
Zou, H., Hastie, T. and Tibshirani, R. (2007) On the "degrees of freedom" of the lasso. *Ann. Statist.*, **35**, 2173–2192.